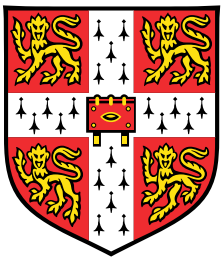
Origin, evolution and global spread of a transmissible cancer clone



Máire Julia Lawlor

Department of Veterinary Medicine

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Newnham College

September 2018

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee

Máire Julia Lawlor
September 2018

Abstract

Origin, evolution, and global spread of a transmissible cancer clone

Máire Julia Lawlor

Clonally transmissible cancers are malignant cellular clones that spread between unrelated individuals by the physical transfer of living cancer cells. These clones can 'metastasise' through populations, having adapted to transmit across external environments and evade the host immune response. Canine Transmissible Venereal Tumour (CTVT) is a sexually transmitted genital tumour that affects dogs and has spread throughout dog populations worldwide. CTVT is the oldest known transmissible cancer and first arose thousands of years ago from the somatic cells of a single, 'founder' dog.

The broad aim of this thesis was to advance understanding of biological features underlying CTVT's continued survival, as well as the evolutionary history that framed the origin of CTVT.

In the first part of this thesis, I profile global tumour diversity using complete mitochondrial genome sequences (mtDNA) from 449 CTVT tumours collected from 39 countries in the first large-scale examination of CTVT clonal diversity. Phylogenetic characterisation of this clonal lineage showed that CTVT has captured mtDNA from transient hosts by horizontal transfer at least five times, defining distinct CTVT clades. Phylogeographic patterns and timings suggest the rapid, recent, multi-route dispersal of CTVT, likely via historic sea routes. Negative selection acts on tumour mtDNA to prevent the accumulation of deleterious mutations and there is evidence for multiple, complex mtDNA recombinations in CTVT. This is the first observed instance of mtDNA recombination in cancer. Enrichment of this data set with almost 200 additional mtDNAs uncovered further horizontal transfer and recombination

events. These findings provide genetic evidence underpinning the importance of functional mitochondria in CTVT evolution.

In the second part of this thesis, by analysing genomic data from ancient and modern canids alongside CTVT whole genomes, I describe the ancestry of the CTVT founder and the spatiotemporal origin of the disease. The CTVT progenitor belonged to a monophyletic lineage of high-latitude dogs that likely originated in North East Asia and dispersed into the Americas alongside people. Using a pair of CTVT biopsies derived from a naturally occurring direct transmission with a known transmission time interval, I estimated the somatic mutation rate in CTVT and inferred the time of CTVT origin. Finally, I make a case for systematically screening cancers with high prevalence in wildlife species for a transmissible cancer etiology.

This work traces the emergence of CTVT, along with detailed geographical and temporal routes of transmissible cancer disease spread over the past two thousand years, and offers new perspectives on the history of dogs as well as key insights into biological mechanisms driving the hitchhiking cancer that has accompanied them around the world.

Acknowledgements

I would like to offer a sincere and heartfelt thanks to my supervisor for welcoming me into her group, for her invaluable support, mentorship and enthusiasm during my PhD and for giving me the opportunity to work on such an incredibly exciting subject. Thank you, Liz!

I am indebted to the many wonderful members of the Transmissible Cancer Group (TCG) for their guidance and support. In particular, the main work of this dissertation would not have been possible without the enormous sampling achievements of Andrea Strakova. I am hugely grateful to Kevin Gori, Young-Mi Kwon, Adrian Baez-Ortega, Tracy Wang, and Maximilian Stammnitz for always providing enthusiasm, humour, inspiration, warm friendship and coffee.

I am extremely grateful to the many collaborators that have enabled the work of the TCG, especially to the veterinary collaborators that have contributed vital samples to this project. I am also indebted to the funding bodies that have generously supported this research, in particular the Wellcome Trust and Newnham College. With immense gratitude, I also thank Laurent Frantz and Greger Larson at the University of Oxford for allowing me to journey through the ancient world of dogs and collaborating on such a fantastic project.

I am more grateful than I can say to all of the people that have made my time in Cambridge so wonderful. Thank you to Aisling for many visits, calls and chats. Thank you to Ana, Ruadhaí, James Gundry, Chris King, Chris Halcrow, Alex, John Riz, Chris Blair, Joe and Alasdair for the happy hours, beer fests and big ones. A heartfelt thanks goes out to all of the wonderful people that passed through 111 King Street and helped me to make a home in Cambridge - Carmen, Andrew, Ali, Paul, Leo, Aoife, Chiara, Sophia, Emma and Andrzej. I owe an extra special thanks to Rox for her unflagging friendship and especially for her care and kindness during the

final days of my write-up.

Thank you to the irrepressible Working Gals: Sinéad, Kate, Gillian and Lisa. I am very lucky to have spent so many years surrounded by such inspiring, driven and brilliant women.

Finally and foremost, a very important thank you to my family and to my partner, Killian, for enduring with me. Agus go háirithe do mo mháthair iontach, Eibhlín.

Go raibh míle maith agaibh go léir!

Preface

The research described in this thesis, unless otherwise stated, was performed at the Department of Veterinary Medicine, University of Cambridge, under the direction of Elizabeth Murchison.

The work undertaken during my doctorate was highly collaborative in nature. Below, I outline the contributions that others have made to the work presented in this thesis. At various points throughout the text, where relevant, I draw further attention to these contributions.

Chapter 2
The CTVT and matched host tissue samples analysed in Chapter 2 (Data Set 1) were contributed by Ilona Airikkala-Otter, Janice Allen, Karen Allum, Leontine Bansse-Issa, Jocelyn Bisson, Artemio Castillo Domracheva, Karina de Castro, Anne Corrigan, Hugh Cran, Jane Crawford, Stephen Cutter, Laura Delgadillo Keenan, Edward Donelan, Ibikunle Faramade, Erika Flores Reynoso, Eleni Fotopoulou, Skye Fruean, Fanny Gallardo-Arrieta, Olga Glebova, Rodrigo Häfelin Manrique, Joaquim Henriques, Natalia Ignatenko, Debbie Koenig, Marta Lanza-Perea, Remo Lobetti, Adriana Lopez Quintana, Thibault Losfelt, Gabriele Marino, Simón Martínez Castañeda, Mayra Martínez-López, Michael Meyer, Berna Nakanwagi, Andrigo De Nardi, Winifred Neunzig, Sally Nixon, Marsden Onsare, Antonio Ortega-Pacheco, Maria Peleteiro, Ruth Pye, John Reece, Jose Rojas Gutierrez, Haleema Sadia, Sheila Schmeling, Olga Shamanova, Richard Ssuna, Audrey Steenland-Smit, Alla Svitich, Ismail Thoya Ngoka, Bogdan Viţălaru, Anna de Vos, Johan de Vos, Oliver Walkinton, Guo-Dong Wang, Ting-Ting Yin, Alvaro Wehrle-Martinez, Mirjam van der Wel, and Sophie Widdowson.

Additional CTVT and matched host tissue samples analysed in Chapter 2 (Data

Set 2) were contributed by Thinlay Bhutia, Cristóbal Briceño, Eric Davis, Adela Espinoza Huerta, Mohammed Fazil, Pagona Gouletsou, Rodrigo Horta, Yaghouba Kane, Cathy King, Ada Krupa, Steven Kruzeniski, Mihran Lazyan, Edward Migneco, Karter Neal, Francisco Pedraza-Ordoñez, Katherine Polak, Jose Rojas Gutierrez, Alan Sherlock, Lester Tapia Martínez, Cristian Torres, Elizabeth Tudor and Sevil Vural.

Chapter 3
The research described in Chapter 3 was partly conducted at the Research Laboratory for Archaeology and History of Art, University of Oxford and supervised by Laurent Frantz and Greger Larson.

Laurent Frantz, Elizabeth Murchison and Greger Larson initiated the project and designed the research.

Archaeological samples from the sites described in Section 3.4.1 were contributed to this study by Angela Perri, Carly Ameen, Jeffrey Blick, Eric Guiry, Anders Hansen, Mikkel-Holger Sinding, Susan Crockford, Morley Eldridge, Vaughan Grimes, John Johnson, Aleksei Kasparov, PAN, Carlos Pereza Lope, Aurelie Manin, Terrance Martin, Kelsey Noack Myers, Mark Omura, Elena Pavolva, Paul Sciulli, Varvara Ivanova, Christopher Widga, and Vladimir Pitulko. Archaeological sample collection was coordinated by Angela Perri, Greger Larson and Keith Dobney with input from Ripan Malhi, Carly Ameen, Ardern Hulme-Beaman, and Kelsey Witt.

Analyses of ancient data was conducted with Laurent Frantz and Evan Irving Pease with input from Elizabeth Murchison and Greger Larson.

Chapter 5
The sample materials analysed in Chapter 5 were contributed by Barbara Halaska and Frances Gulland at The Marine Mammal Institute, California, USA.

Appendix 2
A complete list of publications associated with this thesis, along with references, is included in Appendix 2. While the anglicised version of my name appears on the title page of this thesis, I have used my Irish name (Máire Ní Leathlobhair) in all published work.

Contents

Chapter 1

Introduction to transmissible and non-transmissible cancer

This chapter begins by providing a historical perspective on cancer through a non-exhaustive but representative overview of cancer literature, tracing the major discoveries and findings that have shaped our current understanding of cancer biology. Following this, I introduce an alternative cancer paradigm, transmissible cancers, and give a detailed review of the current knowledge in this field, focusing on the oldest known transmissible cancer, Canine Transmissible Venereal Tumour. The final section of the introduction will summarise key aspects of dog population history and canine genetics relevant to questions explored in Chapter 3. This provides the context for the central questions addressed in this thesis, relating to the emergence and evolution of transmissible cancer in dogs.

1.1 A short history of cancer

1.1.1 An ancient disease

Today, we understand cancer as a family of genetic diseases, all arising from the abnormal clonal outgrowth of a single somatic cell. Different cancers can develop from different cell types and have distinct risk factors and epidemiologies (Stratton et al., 2009). However, all cancer cells converge on a set of acquired capabilities distinguishing them from normal cells. These "hallmarks" include enabling evasion of immune system signalling, recruitment of a permissive environment, angiogenesis, nutrient acquisition, proliferation and invasion of neighbouring and distant tissues (Fouad and Aanei, 2017; Hanahan and Weinberg, 2000). Neoplastic disease is a major cause of mortality in humans, having now surpassed infectious diseases as the second leading cause of death globally (Ferlay et al., 2015). Fatal cases are usually associated with metastasis and subsequent emergence of therapy-resistant disease.

However, cancer is not a recent phenomenon (Faltas, 2011). The earliest historical record describing a disease consistent with cancer is the Edwin Smith Papyrus, one the of oldest medical texts, dating back to circa 2,630 BC (Hajdu, 2016). Medical records from ancient Greece, dating from around the fifth century BC, provide evidence of attempted surgical treatments of malignant growths (David and Zimmerman, 2010). While the earliest purported archaeological evidence for cancer is a 1.7 million year old hominin fossil from South Africa (Odes et al., 2016). Paleoepidemiological studies have found evidence for cancer in ancient Egyptian human remains spanning 3,200–500 BC and individual remains from ancient Germany dating back to 1,800–1,400 BC (Nerlich et al., 2006). Rhabdomyosarcoma, a rare malignancy in modern human populations, was histologically diagnosed in a Chilean mummy, dated between AD 300-600 (Ortner and Aufderheide, 1991).

Cancer susceptibility is an inherent risk of multicellular life. The emergence of complex multicellularity has taken place at least seven times and is one of the major transitions in evolution (Smith and Szathmary, 1997). Multicellular organisms are large, cooperative cell systems where cell-level fitness is sacrificed for organismal fitness. Cancer arises when a single cell in this system ignores the collective imperative and reverts to atavistic unicellular behaviour.

Cancer is believed to arise in almost all metazoans, including dinosaurs (Rothschild et al., 1999), fish (Schlumberger and Lucke, 1948), birds (Effron et al., 1977), beetles (Scharrer and Lochhead, 1950) and molluscs (Barber, 2004). However, the frequency, age of onset and tissue of origin of cancer can vary enormously between species (Misdorp, 1996), as well as between individuals within a single species (Brønden et al., 2010). Interestingly, the observation of naturally occurring tumours in two species of *Hydra* (Domazet-Lošo et al., 2014), a basal metazoan, suggests that cancer has deep evolutionary roots in the metazoan tree. The relationship between organism size, reproductive lifespan, tissue homeostasis and cancer risk across species remains poorly understood. Assuming that all cells have an equal chance of becoming cancerous, we would expect organisms with a larger body mass or that live longer to have an increased risk of developing cancer, constituting a major evolutionary constraint. However, there is no correlation between lifespan or body size and cancer risk, often referred to as 'Peto's paradox' (Caulin and Maley, 2011; Peto et al., 1975).

1.1.2 *Omnis cellula e cellula*

Improvements in modern light microscopes at end of the nineteenth century, and their use in clinical pathology, allowed a look into the microscopic world of cancer and our first real insights into its mechanistic basis. Beginning in 1838, Johannes Müller, a German microscopist, proposed that cancerous growths were composed of cells, at a time when the prevailing view was that cancers were formed from lymph globules or blastema (Haggard and Smith, 1938). He also observed that the cellular forms generally resembled the tissues in which malignant growths arose (Müller, 1838). At about the same time, the surgeon Joseph Récamier, based on observations of secondary deposits in cancer, coined the term "metastasis" from the Greek "methistemi", meaning to displace or change (Récamier, 1829).

Around the time that Darwin proposed a theory of universal common descent, one of Müller's students, Rudolf Virchow, gave a series of lectures suggesting that something similar might be happening in cancer. Credited with coining the phrase "*omnis cellula e cellula*" i.e. "all cells come from cells", he taught that cancer cells descended from normal cells and that 'pathological substitutions' of normal cells with deviant cells lead to cancer (Wagner, 1999).

By the turn of the twentieth century, Boveri, based on observations of aneuploidy in cancer cells, postulated that cancer might arise from a progenitor cell with an anomalous complement of chromosomes (Boveri, 1914). A decade earlier, Von Hansemann had previously described asymmetric mitoses in carcinoma and proposed a link between between defects arising through aberrant cell division and cancer (Hansemann, 1890). Later work would confirm Boveri's finding that specific chromosomal translocations were recurrent in particular cancer types. The best known example being the "Philadelphia chromosome", a translocation between chromosomes 9 and 22 occurring in chronic myeloid leukaemia (Rowley, 1973).

In the middle of the twentieth century, the dual discoveries of deoxyribonucleic acid as the elusive molecular material of inheritance (Avery et al., 1944) and of its molecular structure (Watson and Crick, 1953) catalysed understanding of the genetic pathology underlying cancer. DNA was revealed as the "ultimate target of carcinogenic activation" (Tabin et al., 1982). In 1953, Carl Nordling published his multi-mutation "theory on cancer-inducing mechanism" (Nordling, 1953). In a

statistical study of retinoblastoma, Alfred Knudson proposed a 'two-hit' model of tumourigenesis, whereby two mutational events, likely two recessive mutations in two alleles of the same gene (Knudson, 1973), were sufficient to cause malignancy (Knudson, 1971). In the inherited form of the disease, the understanding was that one of these genetic 'hits' was present since conception in the germline, requiring only a single further hit to promote retinoblastoma. This model was later substantiated by the identification of the biallelic inactivation of *RB1*, the first identified tumour suppressor gene, in retinoblastoma (Murphree and Benedict, 1984).

Genetic analyses comparing transforming varieties of the cancer-causing Rous Sarcoma Virus with non-transforming yet replication-competent varieties showed that the essential transformative factor was an *src* gene (v-*src*), the first identification of an oncogene (Duesberg and Vogt, 1970). Later discovery of c-*src*, cellular DNA homologous to v-*src*, showed that this viral oncogene had a cellular homologue. Importantly, transfection of normal NIH3T3 mouse fibroblasts with genomic DNA from a range of human and animal cancers (Shih et al., 1981) was shown to transform normal cells into cancer cells. This transformation was driven by a number of different cancer-causing genes, mutant versions of normal growth-controlling genes, which were termed proto-oncogenes (Perucho et al., 1981; Pulciani et al., 1982). Shih's canonical study eventually paved the way towards the identification of the first human cancer-causing sequence change: a single base G > T substitution in codon 12 of the *HRAS* gene (Reddy et al., 1982).

These studies largely established the paradigm that has prevailed in cancer research over the past 40 years, focussed on the systematic identification of altered genes (so-called 'cancer genes'), their functional characterisation, and the development of biochemical strategies to target these genes.

1.1.3 Somatic mutation

Somatic mutations accumulate in all cells over time and may be acquired at any stage of the cell lineage, from the zygote to the cancer cell. All cancer cell genomes sampled so far carry somatic mutations. Key somatic changes involved in the evolution of cancer include single nucleotide variants (SNVs), insertions or deletions (indels) of small or large segments of DNA, copy number changes, and structural

rearrangements. These somatic changes accumulate over the course of many cell divisions although not necessarily at a steady rate. In certain circumstances, cells can be subject to violent bursts of mutation; for example, kataegis, localised mutation clusters of SNVs on the same chromosome and chromosomal strand, has been described in breast cancer (Nik-Zainal et al., 2012). Other cancer specific mutational processes include chromoplexy and chromothripsis, both examples of 'all-at-once' chromosome shattering and rearrangement (Forment et al., 2012).

Mutations can be classified based on the role they play in cancer initiation and progression. Driver mutations are those that are causally involved in oncogenesis while deleterious mutations are those that impair cell proliferation or survival Most drivers arise somatically during the lifetime of a patient, although some may be inherited in the germline. The majority of somatic mutations in cancer, however, are selectively neutral 'passenger' mutations that frequently hitchhike alongside driver mutations and have no effect on the fitness of the cell (at least none that are detectable over the lifespan of a cancer). Positive selection acting on somatic cells carrying driver mutations is the prevailing evolutionary pressure in cancer and negative selection is scarcely detectable in human cancer (Martincorena et al., 2017).

1.1.4 Clonal evolution

The principles of cancer evolution, first set out by Cairns and Nowell (Cairns, 1975; Nowell, 1976), follow a Darwinian logic. Classical Darwinian evolution requires genetic variation among the individuals in a population and competition among those individuals for limited resources. The progression from normal to premalignant to malignant cell is an evolutionary process driven by somatic mutation, clonal selection and genetic drift (Stratton et al., 2009). Some somatic mutations can result in phenotypic changes. Selection acting on this genetic and epigenetic variation fuels evolution and adaption in populations of cells. It is important to note that selection takes place within the context of a complex tissue environment, the tumour microenvironment, and consists not only of tumour cells, but also blood vessels, lymph vessels, stromal cells, and immune/inflammatory cells. As individual cells compete with one another for access to host resources, variation that confers a growth or survival advantage will be selected for and drive waves of clonal expansion and tumour progression (Fig. 1.1). Subsequent dissemination of specialised

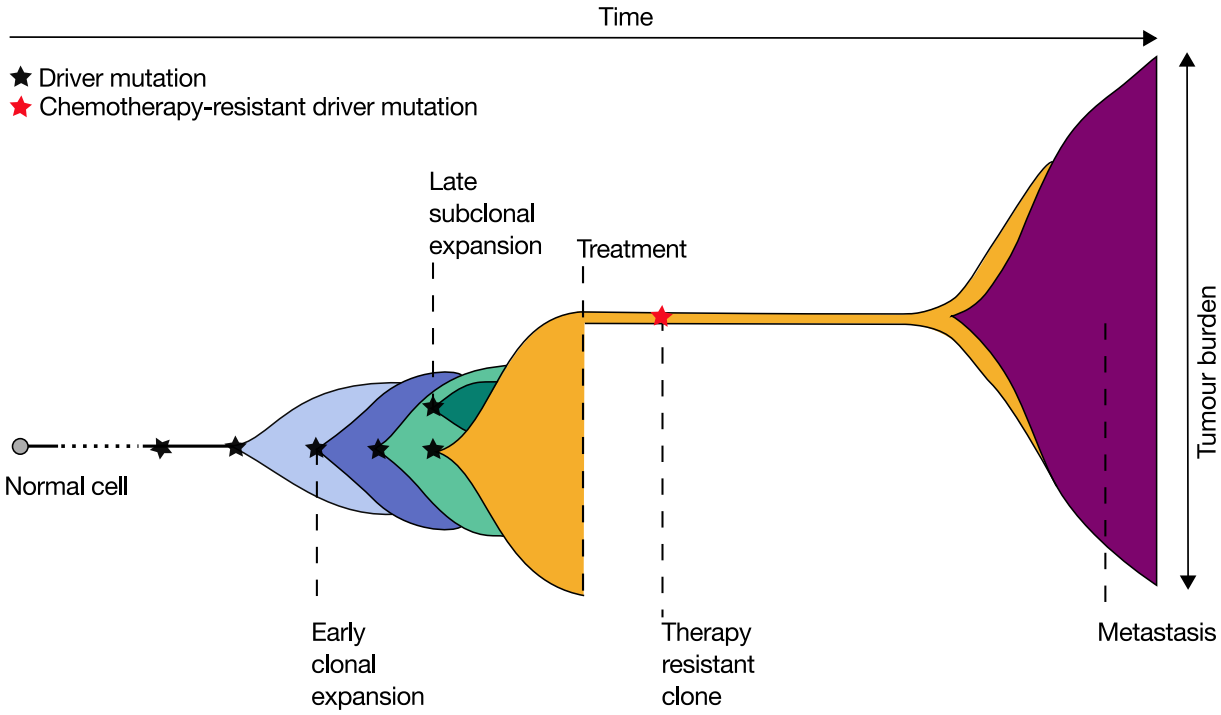


Fig. 1.1. Model of clonal evolution in tumour cell populations. Cancer begins with an initiating driver mutation that confers a selective survival or growth advantage in a normal cell. Drivers promote successive waves of clonal expansion in association with the acquisition of additional mutations.

metastatic tumour cell subpopulations can seed monoclonal or polyclonal outgrowth of secondary tumours (metastases) and disease progression (Fidler, 2003; Gundem et al., 2015).

While the experimental approaches described so far allowed identification of large genomic events (e.g. copy number changes, chromosomal translocations, etc.), the advent of sequencing technologies allowed systematic analysis of the cancer genome at base-pair resolution. Comparison of genetic sequences from a patient's cancer cells with a patient's germline DNA sequences gave the power to observe previously unseen mutational events (Dulbecco, 1986) and promised significant insights into the mechanisms of cancer development.

Early systematic genome interrogations using targeted capillary sequencing identified, what are now, established cancer driver genes including *BRAF* (Davies et al., 2002), *PIK3CA* (Samuels et al., 2004) and *EGFR* (Guillermo Paez et al., 2004; Lynch et al., 2004). The rise of next-generation sequencing methods have allowed increasingly faster and cheaper sequencing of cancer genomes. In particular, massively parallel sequencing has catalysed a technological revolution that has changed the way we understand cancer. Ambitious large-scale initiatives, primarily the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), have sequenced tens of thousands of cancer genomes and matching normal tissues across different tumour types. Recently, even more sophisticated strategies have emerged aimed at understanding how the cancer genome changes over time and space; these strategic advances include multi-region sequencing the same tumour (Gerlinger et al., 2012), 'longitudinal' sequencing of the same tumour at different time points (Ding et al., 2010), single cell sequencing (Navin et al., 2011) and development of cancer organoid cultures (Gao et al., 2014).

1.1.5 Mutational signatures

Recently developed analytical methods, as well as improved sequencing technologies, mean that we can now probe the processes causing somatic mutation by extracting biologically informative 'mutational signatures' from cancer genomes (Baez-Ortega and Gori, 2017). A pivotal study by Alexandrov et al. (2013) found that discrete patterns of single-nucleotide variants and the sequence contexts in which they occurred

in cancer were largely determined by mutagenic processes acting over the lifetime of a cancer cell (Alexandrov et al., 2013). Distinct signatures are generated by endogenous cellular processes, such as defective DNA repair or replication (Alexandrov et al., 2015), and endogenous mutagenic exposures, such as mutagenic enzymes, as well as exogenous mutagen exposures, like ultraviolet (UV) light (Alexandrov et al., 2013) and tobacco carcinogens (Alexandrov et al., 2016). For example, melanomas exhibit far more frequent C>T transitions on the untranscribed compared to transcribed strands resulting from UV induced formation of pyrimidine dimers (Alexandrov et al., 2015). Context-specific signatures driving other somatic changes in cancer, such as genome rearrangements (Nik-Zainal et al., 2016) and copy number changes (Macintyre et al., 2018), have also since been reported.

1.1.6 Cancer and infectious agents

As well as chemical and physical agents, biological agents, such as viruses, bacteria and parasites, can also lead to the cancerous transformation of cells. Following the discovery of a new class of infectious agent, the virus, at the tail end of the nineteenth century (Ivanowski, 1892; Lecoq, 2001), Peyton Rous first showed that solid tumours could be induced in healthy chickens by an infectious "filterable agent", now known as Rous Sarcoma Virus (Rous, 1911). At the time one of Rous's senior colleagues remarked (Becsei-Kilborn, 2010).

> "But my dear fellow, don't you see? They cannot be cancer. You've found their cause!"

Rous was not the first to speculate about the transmissibility of cancer. In 1844, the Italian physician Domenico Antonio Rigoni-Stern observed that cervical cancer was vastly more frequent among married women in Verona than in nuns in the surrounding countryside (Scotto and Bailar, 1969). These observations were later substantiated by the discovery of human papilloma virus (HPV) (Dürst et al., 1983). In humans, infection with viruses such as HPV, Epstein–Barr virus (EBV), hepatitis B virus (HBV), hepatitis C virus (HCV), human T-cell lymphotropic virus (HTLV-1) and Kaposi's associated sarcoma virus (KSHV) is linked to 10-15% of the global cancer burden (Plummer et al., 2016). Viral infection is also linked with cancer incidence and emergence in animals; some notable examples include a polyomavirus associated with the occurrence of neuroglial tumours in raccoons (Curry et al., 2000), first reported in 2013, and Chelonid alphaherpesvirus 5 (ChHV5) linked with fibropapillomatosis

in marine turtles.

Bacteria have also been linked to oncogenesis; *Helicobacter pylori* was the first bacterium to be classified by the World Health Organization as a carcinogen (on the Evaluation of Carcinogenic Risks to Humans, 1994) based on causal association of *H. pylori* infection and gastric cancers. Infection with pathogenic parasites including blood fluke trematodes of the genus *Schistosoma haematobium* and the liver flukes *Clonorchis sinensis* and *Opisthorchis viverrini* is carcinogenic by the International Agency for Research on Cancer (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 2012). However, in all cases, these cancers arise through transformation of the host cells rather than transmission of cancer cells.

So far, I have provided a summary of the landmark genetic and molecular biological studies leading up to the status quo in our current understanding of cancer genetics. In the next section, I will present a challenge to the status quo: transmissible cancers.

1.2 Transmissible Cancers

Under ordinary circumstances, cancer is an individuated phenomenon and continued replication of cancer is normally cut-short by the lifespan of its host. Although "normal" cancer cells can metastasise to different sites within the body they are limited to a single host. Transmissible cancers are rare occurrences whereby cancer cell clones can spread between unrelated individuals, transmitted by the allogeneic transfer of living cancer cells.

Transmissible cancers arise via a favourable combination of specific host and tumour cell traits. Firstly, a cancer-affected individual must shed cancer cells; these cells must be robust enough to transmit across external environments and there must be a route for cell exchange with other individuals in the population. It also requires either that the sloughed neoplastic cells are pre-adapted to evade immune responses and can exploit an allogeneic host or a permissive host environment susceptible to invasion.

Eight transmissible cancer lineages have been described in six different species: one in domestic dogs (Murchison et al., 2014; Murgia et al., 2006; Rebbeck et al., 2009),

two lineages in Tasmanian devils (Pearse and Swift, 2006; Pye et al., 2016) as well as multiple independent lineages in marine bivalves (Metzger et al., 2015, 2016).

Tasmanian devils (*Sarcophilus harrisii*) are the largest living carnivorous marsupials and endemic to mainland Tasmania. This species is affected by two transmissible cancer cell lineages that arose independently about twenty years apart (Table 1.1), known as devil facial tumour 1 (DFT1) (Pearse and Swift, 2006) and devil facial tumour 2 (DFT2) (Pye et al., 2016). Both clones appear as facial and oral tumours with similar gross features, and both diseases are transmitted between devils by biting. DFT1 and DFT2 are genetically and histologically distinct, while karyotype studies suggest that DFT2 arose in a male devil and not the female founder of DFT1 (Deakin et al., 2012; Murchison et al., 2012; Pye et al., 2016). However, intriguingly, there are strong indicators that both DFTs derive from the same tissue type and arose via similar oncogenic processes (Stammnitz et al., 2018). While DFT2 is so far limited to a small number of individuals in south-east Tasmania, DFT1 has resulted in large-scale host population decline and continues to spread (Hawkins et al., 2006).

Five distinct transmissible cancer lineages affect four species of marine bivalves (Metzger et al., 2015, 2016). Disseminated neoplasia, a leukemia-like disease, is transmissible in soft-shell clams (*Mya arenaria*) (Metzger et al., 2015), mussels (*Mytilus trossulus*), golden carpet shell clams (*Polititapes aureus*) (Metzger et al., 2016) and cockles (*Cerastoderma edule*). Analysis of mitochondrial and nuclear markers showed identical tumour genotypes across affected individuals distinct from host genotypes, a hallmark of transmissible cancer, confirming widespread transmission of independent cancer lineages. Two different lineages of transmissible neoplasia, with different morphologies, have been described in common cockles (Metzger et al., 2016), similar to DFT1 and DFT2. Cross-species transmission of cancer cells has also been demonstrated from the pullet shell clam (*Venerupis corrugata*) to a related species in the same region, *P. aureus*. Although the mechanism of transmission is not yet clear it has been suggested that engraftment occurs during filtration of seawater contaminated with transmissible leukemic cells; it is remarkable that clam beds as far as hundreds of miles apart have been colonised by the same cancer clone (Metzger et al., 2015).

Although rare in nature, once established, transmissible cancers can propagate

Table 1.1 Comparison of known transmissible cancers

Species	Species of origin	Cell of origin	Transmission route	Earliest report	Disease Distribution	Survival outcome
Dog	Dog	Myeloid cell	Mating	1810	Worldwide	Rarely fatal
Tasmanian devil	Tasmanian devil	Schwann cell	Biting	1996	Tasmania (mainland)	Fatal
Tasmanian devil	Tasmanian devil	Schwann cell	Biting	2014	Channel Peninsula, Tasmania	Fatal
Soft-shell clam	Soft-shell clam	Hemocyte	Water filtration	1978		
Golden carpet shell clam	Pullet shell clam	Hemocyte	Water filtration	1978	Galician Coast, Spain	Mostly fatal
Common cockle	Common Cockle	Hemocyte	Water filtration	1978	Galician Coast, Spain	Mostly fatal
Bay Mussel	Bay Mussel	Hemocyte	Water filtration	1978	Pacific Northwest Coast, USA & Canada	Mostly fatal

through populations widely and quickly and persist for thousands of years. The discovery of cancer transmission between sessile bivalves shows that physical contact is not necessary for effective cell transmission. Discoveries of multiple transmissible cancer lineages in a single host species suggests that some species may be more susceptible to the development of transmissible neoplasia than others.

The majority of the research carried out during my dissertation has concerned only one of the transmissible cancers: Canine Transmissible Venereal Tumour. In the following section I will explore details of this disease and its life history in depth while elucidating general features of transmissible cancer.

1.2.1 Canine Transmissible Venereal Tumour

1.2.1.1 Overview

Canine Transmissible Venereal Tumour (CTVT) is a clonally transmitted cancer that affects dogs. Mating is the most common route of CTVT transmission and tumours usually manifest on the external genitalia of male and female dogs. This remarkable cancer first arose several thousand years ago in a single founder individual and survived the death of its host by serially colonising allogeneic hosts.

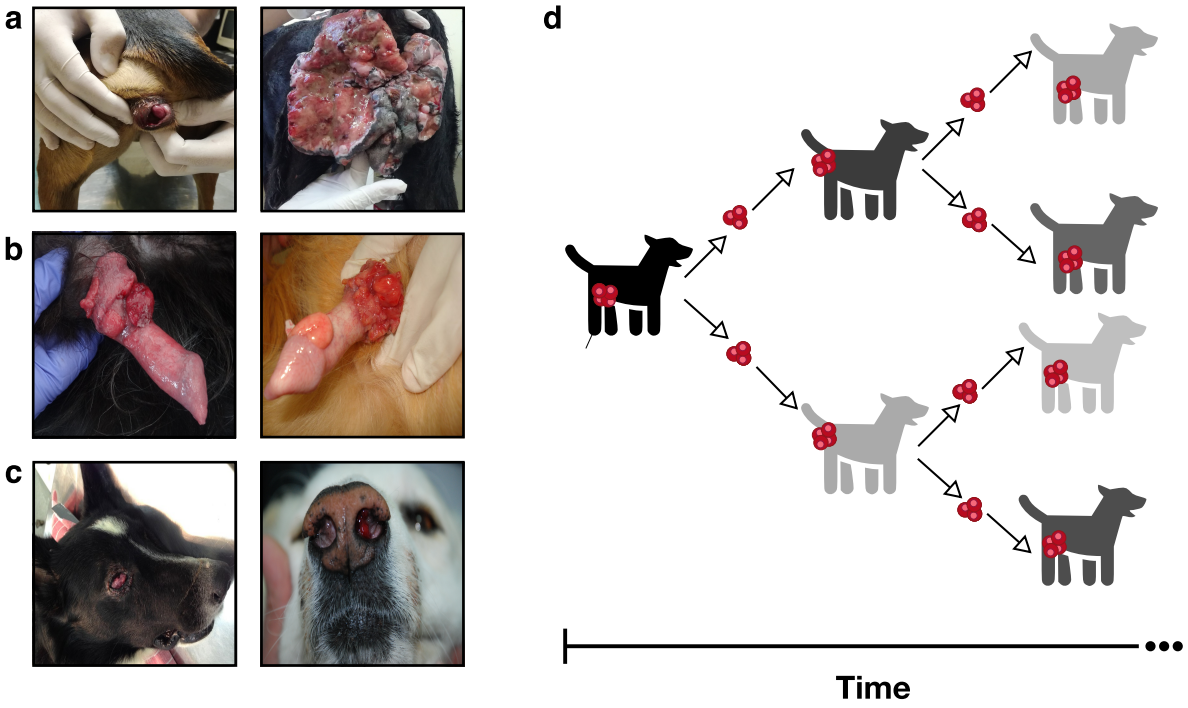


Fig. 1.2. Gross clinical presentation of CTVT in **a.** females **b.** males and, **c.** extra-genital presentation of CTVT in conjunctiva (left) and the nasal cavity (right). **d.** Schematic of CTVT clonal transmission and 'metastasis' through the dog population. Different shaded dogs represent unrelated individuals in the dog population. Red circles correspond to CTVT cells. Images in panels **b** and **c.** were provided by Karina Ferreira de Castro, Natalia Ignatenko, Elizabeth Murchison and Andrea Strakova.

1.2.1.2 Mating and transmission

Cell-transmission between dogs is facilitated by the prolonged copulatory tie in canids as well as injury to the genital mucosa when dogs attempt to separate (Cohen, 1985; Murchison, 2008). Commonly observed extra-genital tumours (Fig. 1.2c) suggests that CTVT cells can also be transmitted between dogs by licking, biting, or sniffing, of tumour-affected areas and during parturition. CTVT has been successfully experimentally transmitted to different canid species, including wolves (Dungern, 1912), jackals (Samso, 1965), coyotes (Cockrill and Beasley, 1979) and red foxes (Wade, 1908).

1.2.1.3 Distribution and prevalence

CTVT has spread around the world, and persists at low clinical prevalence (approximately 1 – 5%) in dog populations across at least 90 countries in all inhabited continents and across a range of diverse altitudes, environments, and climates (Strakova and Murchison, 2014). Disease is generally more prevalent in rural areas and in countries where free-ranging or unmanaged dog populations can act as a reservoir for the disease (Strakova and Murchison, 2014). CTVT largely disappeared from many Western countries, including the United Kingdom, during the twentieth century due to improved management of free-roaming dogs. There is no evidence for gender bias or any variation in susceptibility across different dog breeds (Strakova and Murchison, 2014). Little is known in terms of transmission dynamics; the 'infective dose', or number of cancer cells required to cause infection, has not yet been established. The incubation period, based on reports of naturally occurring transmissions, can range from two weeks to three months (Ajello, 1980; Bellingham Smith and Washbourn, 1898; Locke et al., 1975).

1.2.1.4 Clinical description, diagnosis and treatment

Early-stage tumours commonly appear as small, firm, focal nodules while progressive disease is characterised by multifocal, friable, ulcerated masses with serosanguineous or pure hemorrhagic preputial or vaginal discharge (Fig. 1.2a and b). Tumours are rarely necrotic. Extra-genital localisations of primary tumours have been reported in the conjunctiva (Komnenou et al., 2015), oral and nasal mucosa (Ginel et al., 1995; Rezaei et al., 2016) and skin. Although metastasis is not a characteristic feature of CTVT pathogenesis, cases of metastasis to cutaneous sites, lymphoid tissues, and visceral organs have been reported (Chikweto et al., 2013; Feldman, 1929; Ferreira et al., 2000; Manning and Martin, 1970; Nak et al., 2004; Oduye et al., 1973; Park et al., 2006; Rust, 1949; Sastry et al., 1965; Spence et al., 1978; Varughese et al., 2012).

In most cases, tumours regress following single-drug chemotherapy and tumours are extremely chemosensitive (Brown et al., 1980; Nak et al., 2005). CTVT is routinely treated with vincristine sulphate, a cytotoxic microtubule inhibitor (Amber et al., 1990). Vincristine resistant tumours are usually treated with doxorubicin (Calvert et al., 1982). Other, less commonly used, treatment options include radiotherapy

(Ajello, 1980; Osipov and Golubeva, 1976), immunotherapy (Zander et al., 1980) and surgical debulking.

1.2.1.5 Course of disease

Limited data on the natural clinical course of disease describes long-term persistence in affected hosts and occasional immune-mediated tumour regression, although the frequency with which this occurs remains unclear (Bellingham Smith and Washbourn, 1898; Higgins, 1966; Rust, 1949). Spontaneous regression has been regularly reported in the course of experimental transfer. There is no evidence that CTVT infection is associated with concurrent disease or infection (Strakova and Murchison, 2014) though CTVT is occasionally reported in association with *Leishmania* (Albanese et al., 2002), heartworm (*Dirofilaria immitis*) and ehrlichia (Chikweto et al., 2013).

CTVT cells have a distinct cytological profile (Fig. 1.3). Tumours are composed of round cells with large, central nuclei and prominent nucleoli accompanied by clear cytoplasmic vacuoles and variably prominent intra-tumoural infiltrate of eosinophils. In clinical practice, CTVT diagnosis is based on gross characteristics and cytology, as well as environment history (Duncan and Prasse, 1979; Jackson, 1944). Histologically, CTVT is characterised by diffuse clusters of cells separated by delicate fibrovascular stroma and variable amounts of host immune cell infiltrate (Cohen, 1985).

1.2.1.6 Clonal origin

A disease consistent with CTVT was first described by a British veterinarian, Delabere Blaine, in 1810 (Blaine, 1810). Evidence for clonality was first provided by experimental transplantation of the tumour. In 1876, Mstislav Novinski (Novinski, 1876) carried out the first successful tumour transplantation: serial passage of 'myxosarcoma' between dogs. He believed that successful transfer was based on the inoculation of healthy tissue with 'a living element of the tumour', yet it would be over a century before the clonal origin of CTVT was confirmed. CTVT continued to be widely studied and remained a popular model of tumour transplantation for cancer biologists throughout the late 1800s and early 1900s, before the development of syngeneic mouse and rat systems (Shimkin, 1955). Transmissibility of CTVT was demonstrated consistently by further experimental transfer studies (Karlson and Mann, 1952; Sticker, 1906; Wehr, 1878).

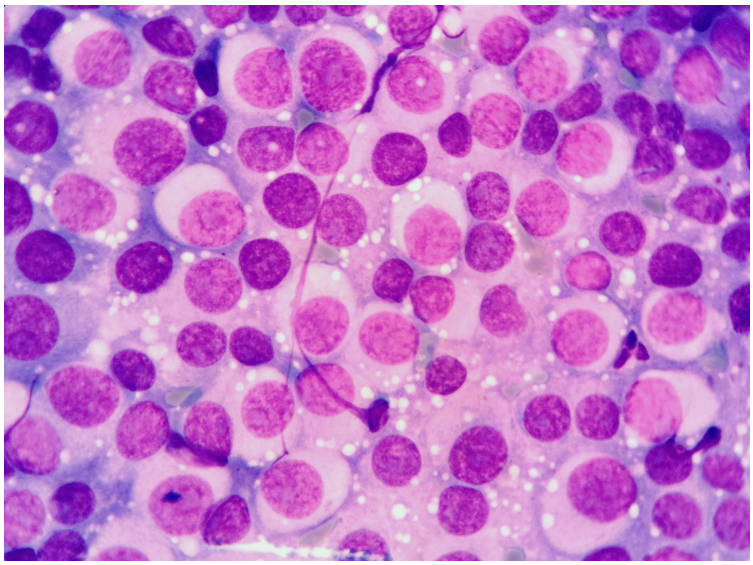


Fig. 1.3. Characteristic CTVT cytology. Prominent nucleoli and cytoplasmic vacuoles can be discerned. Image generated using $\times 100$ objective lens and $\times 10$ ocular lens and reproduced, with permission, from Olga Glebova.

Cytogenetic studies showed that all tumours shared a characteristic rearranged karyotype distinct from the matched host karyotype. The normal karyotype in dogs, wolves, jackals and coyotes is 76 autosomes, all acrocentric, plus metacentric X and Y sex chromosomes while the chromosome number in CTVT is 57-59, at least 16 of which are metacentric (Murray et al., 1969; Oshimura et al., 1973; Weber et al., 1965). Cytogenetically, at least, CTVT no longer resembles a canid.

Further evidence underpinning the clonal origin of CTVT was provided by molecular studies identifying the insertion of a long interspersed nuclear element (LINE) retrotransposon upstream of the *MYC* (c-*myc*) locus found uniquely in CTVT (Katzir et al., 1985). Subsequent analyses of different molecular genetic markers including microsatellites, major histocompatibility (MHC) loci as well as microarray-based comparative genomic hybridization (aCGH) of matched tumour and normal host tissues from naturally occurring tumours collected from five continents (Murgia et al., 2006; Rebbeck et al., 2009) and whole genome sequencing of two CTVT tumours collected in Australia and Brazil (Murchison et al., 2014) confirm that CTVT arose from a single, ancestral somatic cell.

1.2.1.7 The CTVT founder

Early estimates of when CTVT first arose relied on nuclear microsatellite mutation rates in mammalian germlines (Murgia et al., 2006; Rebbeck et al., 2009). These have led to estimates of 250 to 2,500 years (Murgia et al., 2006) and 6,500 to 65,000 years (Rebbeck et al., 2009) since the emergence of the CTVT clone and 47-470 years since the most recent common ancestor (MRCA) of group of globally dispersed tumours. Based on mutation rates estimated from clocklike mutational processes in human medulloblastoma, CTVT is estimated to have arisen between 10,179-12,873 years ago with the most recent common ancestor of sampled tumours existing 460 year ago (Murchison et al., 2014).

CTVT is characterised as a histiocytic tumour and is proposed to be of a macrophage or dendritic cell lineage. Expression of vimentin, lysozyme, alpha-1-antitrypsin, and ACM1 antigen in CTVT tissues supports a histiocytic origin of CTVT (Gimeno et al., 1995; Mozos et al., 1996; Sandusky et al., 1987). The observation of *Leishmania infantum* within CTVT cells (Section 1.2.1.5) also supports a macrophage origin (Carreira et al., 2014; Catone et al., 2003; Marino et al., 2012) as macrophages are the cell type most commonly invaded by *Leishmania* parasites (Liu and Uzonna, 2012).

Analyses of sex chromosome copy numbers in CTVT revealed that CTVT cells contain a single X chromosome and no Y chromosome (Murchison et al., 2014; Thomas et al., 2009). This is consistent with tumour emergence in either a male canid with subsequent somatic loss of the Y chromosome or in a female founder with somatic loss of the X chromosome. Y chromosome loss has previously been associated with highly rearranged cancers (Kujawski et al., 2004).

The first studies looking into the ancestry of the CTVT founder suggested that the disease may have originated in a wolf based on analyses of microsatellite loci, MHC markers (Murgia et al., 2006) and the polymorphic *RPPH1* gene (Rebbeck et al., 2009). However, these studies were based on limited markers. Analysis of polymorphic SNP loci across modern dog and wolf populations indicated that Arctic spitz type dogs, like the Alaskan malamute and Greenland sledge dog, are the closest extant relatives of the dog that first spawned CTVT (Decker et al., 2015; Murchison et al., 2014). Low levels of genetic heterozygosity in CTVT suggest that this individual was relatively inbred and belonged to an isolated, high-kinship population of dogs

Fig. 1.4. Reconstructed phenotype of the extinct CTVT founder dog based on analysis of loci associated with canine phenotypic features reported in Murchison et al. (2014). The founder was likely a medium or large size with a solid black or agouti coat. (Artist, Emma Werner)

(Murchison et al., 2014; Rebbeck et al., 2009). A previous assessment of genetic markers associated with phenotypic traits in CTVT suggested that the founder dog was a medium or large size with an agouti or solid black coat (Murchison et al., 2014). A mosaic of wolflike and doglike alleles were observed at loci associated with domestication (Murchison et al., 2014).

1.2.1.8 The CTVT genome

CTVT has evolved as an asexual mammalian clone without any transfer of nuclear genes between tumours (Rebbeck et al., 2009). Despite thousands of years of mutational onslaught, CTVT appears to have maintained a remarkably stable genome. While cytogenetics shows that the CTVT genome is grossly aneuploid (Section 1.2.1.6), its karyotype is relatively stable across globally distributed tumours suggesting a period of large-scale genomic instability early on in CTVT evolution, during which the genome became rearranged. Whole genome analysis shows that the genome remains largely diploid but with widespread loss-of-heterozygosity (Murchison et al., 2014).

The CTVT genome bears a huge mutational load, even when compared with the most frequently mutated human cancers (Alexandrov et al., 2013), having acquired

between 1 to 2 million somatic single base changes as well as thousands of rearrangements, copy number changes and structural variants (Decker et al., 2015; Murchison et al., 2014). Despite its aneuploidy and massive mutational burden, CTVT appears to have achieved a genome configuration compatible with long term survival (Rebbeck et al., 2009).

Potential early driver mutations have been reported in CTVT, including homozygous loss of *CDKN2A* and *SETD2* as well as rearrangements involving *MYC*, *NEK1* and *ERG*, although the extent to which positive selection acts on these has not yet been explored (Decker et al., 2015; Murchison et al., 2014). The *MYC* rearrangement is common to all tumours suggesting that this was either a germline variant present in the original dog or an early somatic event. If the LINE1 rearrangement was acquired during tumour progression then subsequent *MYC* activation could have played a potential role in the malignant phenotype of CTVT. Given that CTVT is an asexual pathogen, it's likely that many genes in the dog genome are unimportant for its continued survival. At least 646 genes have been inactivated either through homozygous deletion or biallelic inactivating nonsense mutations (Murchison et al., 2014).

The pattern of somatic mutations observed in CTVT can be explained by a combination of three predominant mutational processes or signatures (Murchison et al., 2014). These signatures reflect endogenous mutational processes (COSMIC signatures 1 and 5) and exposure to ultraviolet (UV) light (COSMIC signature 7; Forbes et al., 2015). We would expect cells on the surface of exposed tumours to be exposed to UV radiation; the UV signature in CTVT is a genetic imprint of long-term transmission and suggests that it is these external cells that slough off during physical contact and seed further CTVT tumours.

1.2.1.9 Mitochondrial horizontal transfer in CTVT

Despite the established clonal origin of the CTVT nuclear genome (Murchison et al., 2014; Murgia et al., 2006; Rebbeck et al., 2009), phylogenetic analyses of CTVT mitochondrial sequences showed that CTVT mtDNA clusters in multiple disparate groups (Murgia et al., 2006). It was later suggested that this was the result of periodic capture of mtDNA from transient dog hosts (Rebbeck et al., 2011). Lateral transfer of mitochondrial DNA from host cells to CTVT cells, followed by fixation,

has occurred at least twice in the recent history of CTVT (Fig. 1.5; Rebbeck et al., 2011).

Clonal transmission of CTVT cells over many thousands of years cells may have lead to a Muller's ratchet process in mitochondrial DNA (Charlesworth and Charlesworth, 1997; Muller, 1964): deleterious mutations, that would otherwise be diffused by sexual recombination, "ratchet up" in asexual populations as a result of genetic drift. Understanding the phenomenon of mitochondrial capture from host cells is the driving motivation behind work presented in Chapter 2 and is discussed at length in that introduction to that chapter.

1.2.1.10 Immunology of CTVT

Although not directly relevant to the results presented in this dissertation, the immunology of CTVT warrants some discussion, since one of the central questions motivating the study of transmissible cancer is how these tumours evade the immune system of multiple hosts.

All cancers, transmissible or not, rely on immune evasion strategies. Predominantly these involve down-regulation of MHC, immunoediting and creation of an immunosuppressive microenvironment. A major constraint of cancer transmission in mammals is the adaptive immune system. The major histocompatibility complex is a gene family encoding proteins essential in the regulation of immune response. MHC gene complexes are present in all jawed vertebrate genomes, though they vary greatly across birds, marsupials and eutherian mammals (Cooper and Alder, 2006). Two classes of highly polymorphic genes, MHC class I and MHC class II, enable allorecognition of antigens, uniquely produced in an individual, by host T-cell receptors (Rosenberg and Singer 1992). MHC class I can be expressed by all nucleated mammalian cells and platelets. MHC class II is typically only expressed by antigen presenting cells. This response primarily mediates graft rejection in unrelated individuals.

During active growth, CTVT cells do not express MHC class II and show low level MHC class I expression (Murgia et al., 2006). Downregulation of MHC in CTVT may result in reduced antigen presentation, enabling the tumour to survive hostile host environments and evade immune system signalling (Murgia et al., 2006). MHC

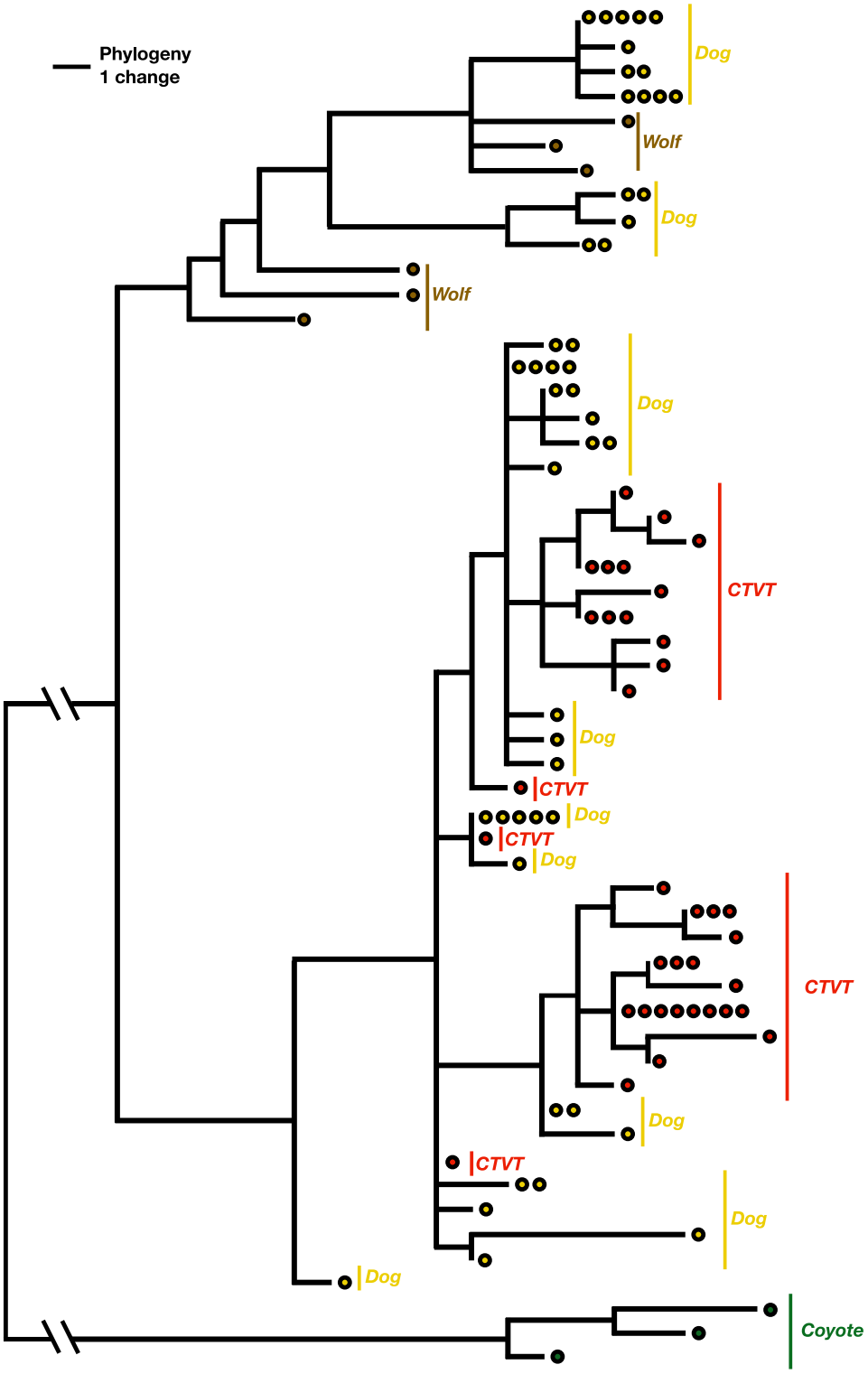


Fig. 1.5. A maximum parsimony phylogeny of mitochondrial sequences from dogs, wolves, and 37 CTVT samples collected across four continents. Individual samples are represented by coloured dots. Scattered phylogenetic placement of CTVT samples suggested that CTVT lineages had periodically acquired the mitochondria of their hosts. Figure adapted from Rebbeck et al. (2011). RightsLink License Number 4431400002345.

expression is restored during spontaneous tumour regression exposing CTVT as a foreign tissue graft (Cohen et al., 1984; Hsiao et al., 2008; Murgia et al., 2006; Yang and Jones, 1973). CTVT can also trigger a humoral immune response and antibodies recognising CTVT antigens can be detected in the sera of CTVT-affected animals (Cohen and Steel, 1972; Epstein and Bennett, 1974; McKenna et al., 2010).

CTVT has been experimentally transmitted as a xenograft in immunodeficient mouse models (Harmelin et al., 2001; Holmes, 1981; Stubbs and Furth, 1934) while attempts to transmit CTVT to immunocompetent rabbits, rats, guinea pigs (Wade, 1908), mice, cats, hamsters, opossums (Cockrill and Beasley, 1979) have been summarily unsuccessful. Although wolves, coyotes and golden jackals can mate successfully with dogs (Galov et al., 2015; Randi et al., 2014) and experimental transmission of CTVT has been carried out successfully in these species (Section 1.2.1.2, naturally occurring cases of CTVT have only been reported in dogs. This could suggest limited interbreeding with wild canids or the existence of some species-specific restriction or risk factors, perhaps signalling between tumour and host cells is essential for the tumour to be tolerated; beyond a certain evolutionary distance, this signalling fails and prevents engraftment.

Tumour progression in CTVT is typically described in terms of three phases: a progressive phase of initial growth (P phase), a stationary phase of stable growth (S phase) and regression (R phase; Epstein et al., 1964). Transition from the "S phase" to regression is characterised by an increase in lymphocyte infiltration, upregulation of MHC (Siddle and Kaufman, 2013) and release of the inflammatory cytokine, interferon gamma (IFN-gamma).

Aggressive tumour growth and metastasis, along with the absence of tumour-infiltrating lymphocytes, has been observed in immunosuppressed dogs and newborn pups infected with CTVT (Yang and Jones, 1973). Pups born to females in which the tumour had spontaneously regressed were able to limit the growth of the injected tumour which could be explained by transfer of immunity (Yang and Jones, 1973). In some cases, dogs that have undergone spontaneous regression are reported to demonstrate immunity to further tumour inoculation (Cohen, 1985).

1.2.2 Cancer cell transmission in humans

There are no known cases of naturally occurring horizontal transmission of cancer between more than two humans, however, there have been some exceptional cases involving transmission between two humans. One of the only viable routes for cancer cell transmission in humans emerges during gestation and cases include transplacental transmission of lymphoma, acute leukaemia, melanoma and carcinoma from mother to foetus (Bolze et al., 2013; Tolar and Neglia, 2003), transfer of leukaemia *in utero* between monozygotic twins (Tolar and Neglia, 2003) and choriocarcinoma, whereby embryonic trophoblast cells derived from the fetus develop in the mother. In one case of maternal-fetal transmission, selective loss of MHC alleles not inherited by the infant was observed in maternally-derived tumour cells transmitted to the infant (Isoda et al., 2009). This underlines that MHC-mediated rejection is an important barrier to cancer cell transmission and that cancer cells transmitted between mother and fetus are more likely to be tolerated because they are semi-allogeneic in their new host.

An unexpected route of cancer cell transmission arises during organ transplantation. Immunosuppression of organ, bone marrow, and blood stem cell transplant recipients has also lead to transmission and development of a range of different donor-derived cancers (Braun-Parvez et al., 2010; Kauffman et al., 2002; Sala-Torra et al., 2006; Strauss and Thomas, 2010). Strikingly, neoplastic cells derived from a dwarf tapeworm, *Hymenolepis nana*, were observed in lung and lymph node biopsies taken from an immunosuppressed patient infected with human immunodeficiency virus (HIV) (Muehlenbachs et al., 2015).

Cases of cancer transmission outside the context of immune-suppression or relatedness between donor and host are extremely rare. Documented cases include the accidental transfer of a malignant histiocytoma from a patient to a genetically unrelated (and unlucky) surgeon (Gärtner et al., 1996), accidental needle-stick transmission of colonic adenocarcinoma to a laboratory worker with no history of immune deficiency (Gugel and Sanders, 1986) and transfer of breast carcinoma to a medical student following accidental puncture with a syringe. The student passed away one year after the incident as a result of advanced metastatic disease (Lecène and Lacassagne, 1926). Chester Southam also famously inoculated inmates at Ohio State Penitentiary with cancer cells (Moore et al., 1957).

Overall, the occurrence of cancer cell transmission in humans is rare and the continuous propagation of a clonal cancer lineage in our species remains unlikely. However, it has been suggested that some human cancers might have the potential to become clonally transmissible between individuals infected with HIV (Metzger et al., 2016). Cancers in humans can render themselves immunologically silent through down-regulation of MHC (Algarra et al., 1999) and perhaps these MHC-negative tumours might be able to thrive in immunocompetent individuals.

Limitless replicative potential is hallmark of cancer (Hanahan and Weinberg, 2000) and, given an appropriate environment and nutrients, cancers have the potential for biological immortality. The first human cells immortalised *in vitro* (Gey et al., 1952), known now as HeLa cells, derived from cervical carcinoma cells biopsied, without informed consent, from Henrietta Lacks in 1951 (Skloot, 2017). HeLa cells constitute the oldest human cancer cell lineage and share an interesting feature with CTVT cells as another highly rearranged, but genetically stable, immortal cell line.

1.2.3 Emergence of transmissible cancers in other animal species

While CTVT, DFT1 and DFT2 are the only naturally occurring transmissible cancers identified in mammals, clonal transmission of cancer cells in animal laboratory populations has been reported. Direct transmission of reticulum cell sarcoma was described in a laboratory population of golden hamsters (*Mesocricetus auratus*) (Brindley and Banfield, 1961) and demonstrated by karyotype analysis (Cooper et al., 1964). Examination of tumour cells through electron microscopy and tissue cell culture as well as failed passage of cell-free material between animals showed that the disease vector was unlikely to be a viral particle (Cooper et al., 1964). Cell transmission between individuals took place by subcutaneous implantation and during contact exposure, most likely via mosquito bite and as a result of cannibalism (Banfield et al., 1965). Various mouse transplantable cancer cells can be propagated through unrelated mouse hosts, including S180 (sarcoma cell line), L5178Y (lymphoma cell line), and EL4 (lymphoma cell line; Patt and Straube, 1956).

Transmission of disseminated neoplasia has been suggested in other species, including bay mussels (*Mytilus edulis*; Elston et al., 1988) and European flat oysters *Ostrea edulis* (Barber, 2004); these might be caused by malignant clonal cell lineages.

Putative transmissible cancers, with unknown etiology, have been described in newts (*Triton alpestris*; Champy and Champy, 1935) and climbing perch (*Anabas scandens*) as well as transplantable tumours in bullhead catfish (*Ameiurus nebulosus*), axolotls (*Siredon mexicanum*) and leopard frogs (*Rana pipiens*; Schlumberger, 1957). Cancer epidemics with unconfirmed aetiology have also been observed in California sea lions and green turtles (*Chelonia mydas*; McAloose and Newton, 2009).

The incidence of transmissible cancer outbreaks in nature could be underestimated as no attempt has been made to systematically screen available wildlife data sets for the presence of clonally transmitted malignant cells. Failure to recognise contagious cancers may be because outbreaks are self-limiting in local populations, having been eliminated from the population due to their detrimental impact on host fitness or may have resulted in species extinction outright (Ujvari et al., 2016). This would be especially true of highly virulent outbreaks. Furthermore, cancer prevalence data relating to cancer in mammals, and in particular birds and reptiles, is limited, scattered and frequently challenging to access (McAloose and Newton, 2009). It is possible that, for these reasons, we have underestimated the frequency with which contagious cancers can arise.

1.3 Dog origins and genetics

A major part of Chapter 3 is framed in terms of dog population history and, for this reason, I include here a brief discussion of the relationship of dogs to wild canids and overview their origin and domestication.

Canidae, a family of enigmatic terrestrial carnivores, is formed by 36 extant canid species including wolves, jackals and coyotes (Wayne et al., 1997). Dogs (*Canis lupus familiaris*) were the first domesticated species of plant or animal, and the only animal species to be domesticated years before the advent of settled agriculture (Larson et al., 2012). Modern global dog populations trace their ancestry to a single population of extinct late Pleistocene grey wolves (*Canis lupus*) that expanded from Beringia, the land bridge joining Alaska and Eurasia that is now submerged, ~20,000 years ago (Fan et al., 2016; Loog et al., 2018). However, the geographic and temporal origin of the dog, along with the number of independent wolf domestications involved, remains hotly contested and debate roils within the fields of canine archaeology and

genetics (Callaway, 2013; Grimm, 2015; Lallensack, 2017).

The earliest unequivocal dog remains appear in the Middle East around 12,000 BP (Dayan, 1994), in Europe at 15,000 BP (Pionnier-Capitan et al., 2011), in North America up to 10,190 cal BP (Perri et al., 2018) and in Africa between 6,300-5,600 BP (Mitchell, 2015). Recent genomic estimates have placed the timing of wolf-dog divergence somewhere between 11,000-16,000 BP (Axelsson et al., 2013; Freedman et al., 2014; Wang et al., 2013). However, a number of contested paleontological reports point towards the presence of "incipient" dogs up to 40,000 BP, before the peak of the Last Glacial Maximum (LGM) across sites in Europe (Belgium, Germany, Czech Republic) and Siberia (Camarós et al., 2016; Germonpré et al., 2015, 2012, 2009; Ovodov et al., 2011). MtDNA analysis of a putative dog specimen dating to 33,000 years from the Altai mountains in central Asia ago support an earlier domestication window (Druzhkova et al., 2013). A molecular timescale of dog and wolf population history, recalibrated using an ancient Siberian wolf genome, places their divergence ~27,000-40,000 years ago (Skoglund et al., 2015).

A single domestication origin has been suggested in Central Asia (Shannon et al., 2015), East Asia (Savolainen et al., 2002; Wang et al., 2016), Europe (Thalmann et al., 2013) and the Middle East (Vonholdt et al., 2010). Recently, a consilient dual domestication origin was suggested in East Asia and Western Eurasia based on combined genetic and archaeological results (Frantz et al. 2016; see Table 1.2 for a brief summary of the major dog domestication studies). The process by which wolves were domesticated, including the specific selected traits and how these were selected, is not well characterised. Given that wolves are large carnivorous predators it seems likely that scavenging wolf commensals initiated a self-domestication process, taking advantage of feeding ecologies provided by human activity and carving out a niche in which the modern dog could emerge (Larson et al., 2012; Pitulko and Kasparov, 2017; Zeder, 2012).

The development of the modern domestic dog is associated with two population bottlenecks: one following domestication and the divergence of wolves and dogs and a far more recent, tight bottleneck following breed creation (Karlsson and Lindblad-Toh, 2008). In early dogs, gene flow between dog and wolf populations was likely an important source of variation for artificial selection (Vilà et al., 1997). Nowadays,

Table 1.2 Overview of dog domestication studies
ND, not determined, signifies either that studies did not propose a geographic origin for domestication, did not propose a temporal origin for domestication, or did not comment on the number of domestication events involved. Ybp, years before present.

Domestication centre	Dog time-of-origin	Domestication events	Data	Publication
ND	<135,000 ybp	ND	partial mtDNA	Vilà et al. (1997)
East Asia	~15,000 ybp	ND	complete mtDNA	Savolainen et al. (2002)
East Asia	<16,300 ybp	1	complete mtDNA	Pang et al. (2009)
Middle East	ND			Vonholdt et al. (2010)
Europe	18,800-32,100 ybp	1	complete mtDNA (ancient & modern)	Thalmann et al. (2013)
ND	11,000-16,000 ybp	1	nuclear DNA	Freedman et al. (2014)
Central Asia	<15,000 ybp	1	nuclear, Y & MT markers	Shannon et al. (2015)
East Asia (southern)	<33,000 ybp	1	nuclear DNA	Wang et al. (2016)
East Asia; Western Eurasia	20,000-60,000 ybp	2	nuclear DNA (ancient & modern)	Frantz et al. (2016)
ND	20,000-40,000 ybp	1	nuclear DNA (ancient & modern)	Botigué et al. (2017)

there are almost 400 modern dog breeds, most of these resulting from a program of intense artificial selection, beginning in the nineteenth century (Parker et al., 2004). However, the majority of dogs are free-roaming village dogs, a geographically widespread and genetically diverse population (Shannon et al., 2015).

1.4 Thesis structure and rationale

This thesis has the following plan. Chapters 2 to 4 describe experimental projects carried out during the course of my dissertation and the results obtained. The first part of Chapter 2 describes the analysis of mitochondrial genomes from a global population of CTVTs; the second part of Chapter 2 describes analysis of an extended set of mtDNAs. I used this data to determine the frequency and timing of mtDNA horizontal transfer in CTVT as well as the potential contribution of selection and recombination to CTVT mtDNA evolution.

Chapter 3 reconstructs the genetic profile of the 'founder dog' in which CTVT first arose along with providing a detailed evolutionary history of the early dog population to which the founder belonged. This chapter also provides a CTVT time-of-origin estimate based on mutation count data from a case of direct CTVT transmission.

Chapter 4 proposes that an infectious aetiology, namely transmissible cancer, could explain wildlife cancer epidemics and specifically tests the hypothesis that urogenital carcinoma in California sea lions is clonally transmitted.

Chapter 5 summarises the key findings and implications of this thesis, which follow from the explicit research questions framed in each chapter, and suggests future directions of research arising from this work.

At the beginning of each chapter I present a focused introduction, providing background details relevant to work presented therein. I include a Discussion section at the end of each chapter, in which I summarise my work and results, as well as the limitations of this work, and future research directions.

Chapter 2

Exploring the mitochondrial genetic diversity and evolution of globally distributed canine transmissible tumours

2.1 Chapter abstract

Mitochondria play a crucial role in cellular pathways frequently disrupted in cancer cells, including energy metabolism and apoptosis. A report of multiple mtDNA horizontal transfers from host to CTVT cells suggested that, in particular, mitochondria might play an important role in CTVT biology and evolution (Section 1.2.1.9). Furthermore, distinct CTVT mitochondrial haplotypes arising as a result of mtDNA horizontal transfer provide a unique means of tracing routes of CTVT phylogeographic spread along with historical global dispersals of dogs.

In the first part of this chapter, I analyse complete mitochondrial genome sequences (mtDNAs) from 449 CTVT tumours collected from 39 countries around the world. This analysis revealed that CTVT has captured mtDNA from its transient hosts at least five times over the last two thousand years. Further analysis of an extended data set of 640 CTVT samples uncovered a total of twenty horizontal transfer events as well as frequent uptake of host mtDNA belonging to a specific haplogroup.

Somatic mtDNA mutations have been shown to accumulate with age (Wallace, 2010). This may be expected to lead to a high mitochondrial DNA mutation load in CTVT, similar to the nuclear genome, given the opportunity for mutations to accumulate since its ancient origin. However, CTVTs persistence in spite of this suggests that negative selection, acting to remove deleterious mutations, could be a major force in CTVT evolution. This chapter explored explicitly whether negative selection operates on CTVT mitochondrial genomes and compared this with the effects of selection in

short-lived human cancers. This analysis demonstrated the activity of negative selection operating to curb the accumulation of deleterious mutations in CTVT mtDNA.

Lastly, this chapter exposes an unprecedented genetic mechanism in cancer, mtDNA recombination, which may be a widespread DNA repair mechanism although challenging to detect in cancers that remain in one host. Altogether, these data present some new perspectives on recent dog migration history, as well as novel insights into the evolutionary processes driving the cancer that has doggedly followed them around the world.

2.2 Publications associated with this chapter

Results from analyses performed in this chapter (Data Set 1 only; Section 2.6) were published as a Short Report in *eLife* on May 17th 2016, of which I am co-first author along with Andrea Strakova (Strakova et al., 2018; see Appendix 2 for the full reference). A manuscript including the main results from analysis of Data Set 2 (Section 2.7) is currently in preparation (Strakova et al., 2018).

2.3 Introduction

This introduction briefly summarises aspects of mitochondrial biology relevant to the research questions addressed in this chapter. I begin by considering the origin of mitochondria along with their structure and function in eukaryotic cells. I introduce some important features of mitochondrial genetics, including how somatic mitochondrial mutations arise and selection and the role of mitochondrial mutation in cancer. I then review recent findings regarding mitochondrial dynamics, specifically horizontal transfer, recombination and mitochondrial-nuclear genome fusions. I conclude by setting out the central questions of this chapter.

2.3.1 Mitochondrial structure and function

Mitochondria are double membrane bound organelles found in almost all eukaryotic cells (Cavalier-Smith, 1987). These intracellular organelles are the direct descendents of a primitive proteobacterium, related to α-Proteobacteria, that became associated with a prokaryotic cell through a fateful endosymbiosis approximately 2 billion years ago (Margulis, 1970).

Since this endosymbiotic origin, mitochondria have co-evolved with nuclear DNA and become indispensable for eukaryotic life. Mitochondria are the primary generators of adenosine triphosphate (ATP), the energy currency of the cell, through the process of oxidative phosphorylation (OXPHOS) and are critical to the regulation of other cellular functions including apoptotic cell death and calcium signalling (Tait and Green, 2010). Individuals cells can contain 10^3-10^4 mitochondria with the number of mitochondria varying by cell type and higher mitochondrial copy number observed in conjunction with greater metabolic demands (Robin and Wong, 1988; Wai et al., 2010). Mitochondria can replicate, even in non-dividing cells, semi-autonomously of nuclear DNA (Westermann, 2010) and are frequently fusing and dividing. These organellar dynamics determine their structure; in some cell types, mitochondria can fuse to form a networked mitochondrial complex (Rafelski, 2013). Mitochondria are usually assumed to be clonally inherited. For this reason, mitochondrial genotypes are popular markers in forensics and in reconstructing population histories.

2.3.2 The mammalian mitochondrial genome

The hypothesis for a symbiotic origin was supported by the discovery that mitochondria contain their own genome. In the course of mitonuclear coevolution, most genes from the ancestral mitochondrion have relocated to the nucleus (Kleine et al., 2009). The mitochondrial (MT) genome is a circular molecule, roughly 14-20 kilobase (kb) in animals. In humans and dogs, this compact genome can be divided into a large coding region, encoding 13 proteins, 22 transfer RNAs (tRNAs) and 2 ribosomal RNAs (rRNAs) and a short, non-coding region, called the displacement loop (D-loop) or control region (Kim et al., 2012). Mitochondria lack introns and, apart from the D-loop, intergenic sequences are limited to a few bases.

The MT genome is composed of two complementary mitochondrial DNA (mtDNA) strands: a heavy and a light strand (referring to molecular weight). There is a distinct asymmetry between mitochondrial strands in terms of gene distribution and nucleotide content. The heavy strand is guanine-rich and encodes for 12 out of the 13 proteins, as well as most tRNAas and rRNAs; only one protein-coding mitochondrial gene, *ND6*, is transcribed from the corresponding cytosine-rich light strand

(Andrews et al., 1999). According to the 'strand-displacement' model of mtDNA replication, transcription within the non-coding mtDNA D-loop initiates replication and proceeds clockwise from the origin of heavy-strand replication (OH) until the origin of light-strand replication (OL) is exposed; this allows light-strand synthesis to proceed clockwise until the entire molecule is copied. The mtDNA genetic code is not the same as that of the nuclear genome. Importantly, in vertebrate mtDNA, UGA is not a stop codon (as in the nuclear genome) and the two arginine triplets, AGA and AGG, are recoded as stop codons (Stewart and Chinnery, 2015).

2.3.2.1 Dog mtDNA

The complete dog mitochondrial genome sequence was published in 1998 and derived from a native Korean dog, the Sapsaree (Kim et al., 1998). Dog mtDNA comprises 16,727 base pairs and the structure of the genome is the same as that in humans (outlined in Section 2.3.2). Dog mitochondrial haplotypes can be divided into six main phylogenetic groups (clades A, B, C, D, E, and F; Fig. 2.1) with almost all modern dog haplotypes belonging to clades A, B and C (Savolainen et al., 2002; Vilà et al., 1997). Clade D, E, and F haplotypes are rare and have a strong geographic signal. It has been suggested that these rarer haplotypes derive from localised hybridisation with wolf populations (Savolainen et al., 2002). Interestingly, dogs have accumulated non-synonymous mutations in mitochondrial genes at a faster rate than wolves. It has been suggested that relaxation of the selective constraint on mitochondria in dogs occurred as a result of domestication (Björnerfeldt et al., 2006).

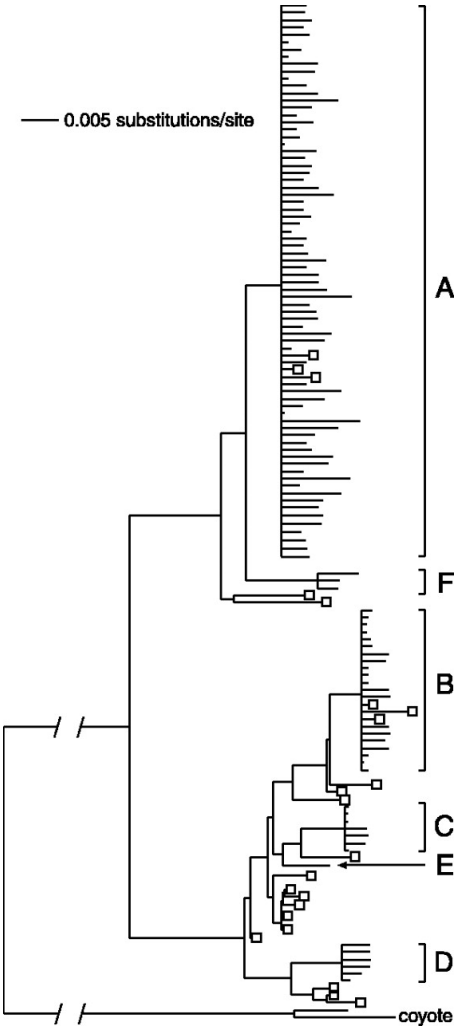


Fig. 2.1. Phylogeny of modern dog (unlabelled) and wolf (open squares) mtDNA haplotypes. Dog sequences were assigned into six phylogenetic groups (clades A, B, C, D, E and F). Branch lengths are according to the indicated scale; the branch leading to the outgroup (coyote) was reduced by 50%. Reproduced from Savolainen et al. (2002). RightsLink License Number 4410730872207.

2.3.3 Mutation and selection

Somatic mutations can arise in the mitochondrial genome just as they do in the nuclear genome and with each cell cycle the replication MT genome can acquire de novo mutations. In mtDNA, somatic variants originate on a single molecule in a single organelle in a single cell (Stewart and Chinnery, 2015) resulting in heteroplasmy. Heteroplasmy

arises when multiple distinct mtDNA populations coexist within the mitochondrion, the cell, or the individual. Clonal expansion of heteroplasmic mtDNA mutations can lead to fixation or homoplasmy, so that the mutation now affects all molecules (Fig. 2.2c.).

In mammals, the rate of mutation in the mitochondrial genome is higher than the in nuclear genome (Lynch et al., 2006). A number of important factors are likely to underlie this elevated mutation rate including: (*i*) mitochondria generate reactive oxygen species (ROS) as a byproduct of respiration; oxidative damage plays a known role in mitochondrial mutagenesis, (*ii*) mtDNA lack protective histones, and (*iii*) mtDNA damage repair is relatively inefficient compared with nuclear DNA repair (Stewart and Chinnery, 2015). This could be, in part, because high mitochondrial copy number has a buffering effect against the deleterious effects of stochastic mutation.

Random genetic drift (Coller et al., 2001), host-beneficial or selfish selection can result in the proliferation of mutant mtDNAs. For example, smaller mitochondria carrying pathogenic deletions have a selfish replicative advantage and have been observed to preferentially accumulate over time in animal and human models (Clark et al., 2012; Diaz et al., 2002). Purifying or negative selection can act to eliminate defective variant mtDNAs. A key example are the mitochondrial bottlenecks that occur in the developing germline, thought to be responsible for the maintenance of homoplasmy and purifying selection of highly deleterious mutations during germline transmission (Stewart et al., 2008). Lastly, balancing selection can lead to a state of stable heteroplasmy (Ma et al., 2014). It is also worth noting that patterns of selection can operate at multiple levels: between mitochondrial genomes, between mitochondria within cells or between cells (Fig. 2.2) and that selection at different levels can lead to genetic conflicts.

2.3.4 Mitochondria and cancer

While mutations in nuclear cancer genomes have been studied extensively, the role of mitochondrial DNA mutations in cancer is not yet completely understood, although somatic mutations in mtDNA have for a long time been suspected to contribute to cancer progression (Chatterjee et al., 2006; Ohta, 2006; Yuan et al., 2017; Zong et al.,

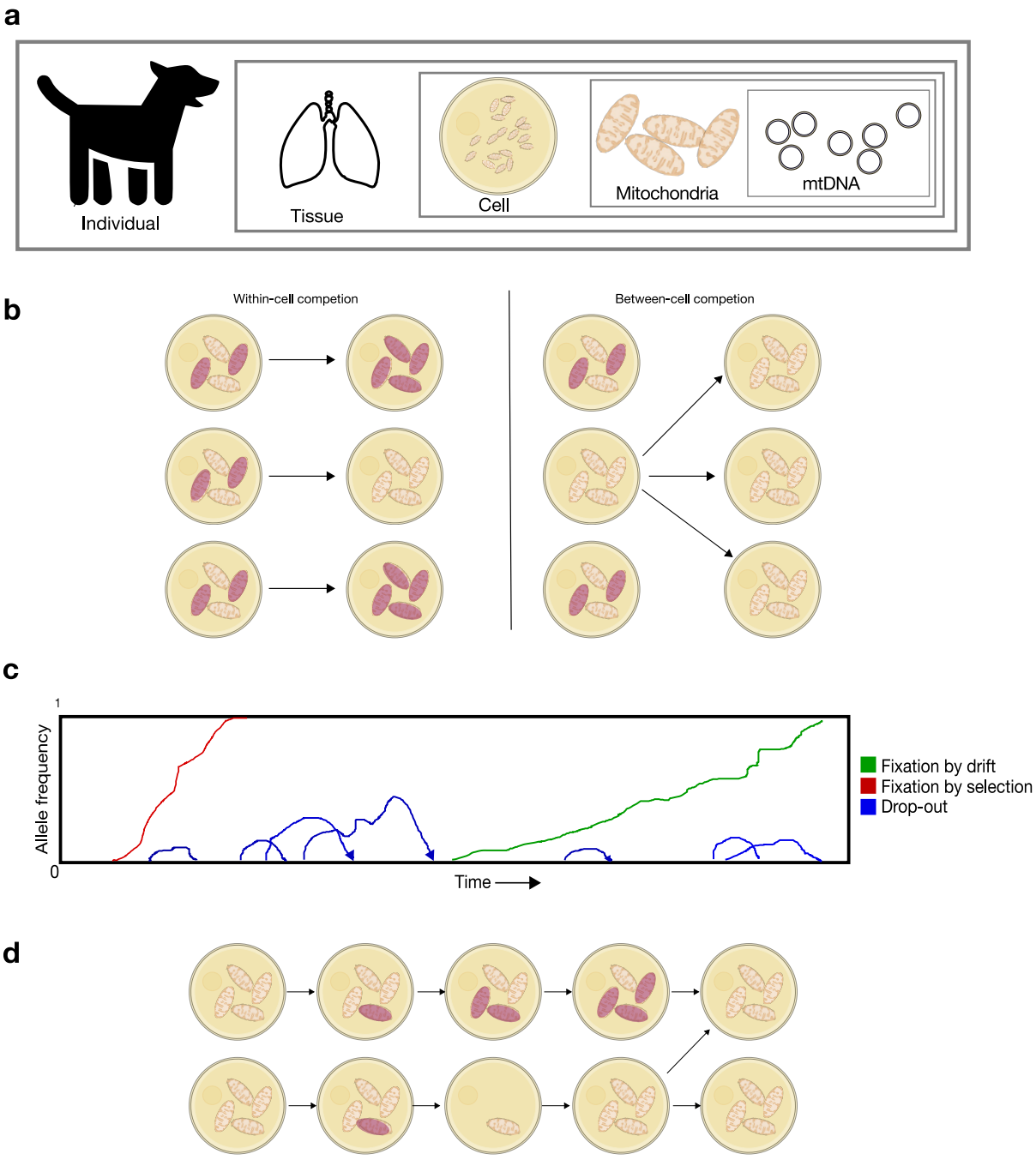


Fig. 2.2. Mutation and selection in mitochondria.
a. Levels of selection within an individual **b.** Multiple levels of selection: between cells, between mitochondria within cells and between mtDNA within mitochondria. **c.** Fixation of mtDNA mutations **d.** Between-cell selection acting against gene-disrupting mutations effective only with MT bottlenecks

2016). It has been hypothesised that because of the critical role of mitochondria in energy metabolism and apoptosis, mitochondrial mutations can contribute to cancer development. Notably, cancer cells are characterised by an altered energy metabolism (Hanahan and Weinberg, 2000) and early interest in the role that mitochondria might play in cancer stemmed from this observation. Otto Warburg famously observed that cancer cells metabolise glucose in a way that differs from normal differentiated cells, the so-called "Warburg effect", and primarily rely on aerobic glycolysis for ATP synthesis rather than mitochondrial OXPHOS (Vander Heiden et al., 2009; Warburg et al., 1924).

Despite this, comprehensive analysis of somatic mtDNA variants in 1,675 tumours (Ju et al., 2014) found that somatic mutations in the mitochondrial genome accumulate largely neutrally as a result of an endogenous mechanism linked with replication. The same study did report, however, that protein-truncating mtDNA mutations in human cancer cells undergo negative selection indicating the need for a sufficient number of functioning mitochondria in order to maintain fitness of the cancer cell (Ju et al., 2014).

2.3.5 Horizontal transfer

The extent of mitochondrial horizontal transfer between mammalian cells in different tissues and contexts is currently unknown. Lateral transfer of mtDNA has been observed in many different human and animal cell culture systems (Berridge et al., 2015; Cho et al., 2012; Rustom et al., 2004; Spees et al., 2006) as well as in normal and cancer mouse cells *in vivo* (Berridge et al., 2015, 2016; Dong et al., 2017; Hayakawa et al., 2016; Spees et al., 2006; Tan et al., 2015). There is also evidence that many cell types, including cancer cells, can receive mitochondria (endothelial, epithelial, macrophage; Berridge et al., 2016) and that intercellular mitochondrial trafficking can take place between the same or different cell types.

Mitochondria are sensitive to changes in the cellular environment, and are among the first organelles to be damaged under stressful conditions, like low oxygen levels or rapid reperfusion. These damaged mitochondria can quickly become drivers of apoptosis. It has been suggested that horizontal transfer can act to rescue cell function and may be a response to metabolic stress or other changes within the cell

environment. For example, tumour cell uptake of donor DNA from host cells in the microenvironment has been shown to rescue mitochondrial respiratory function in tumour cells with severely damaged mtDNA and to promote tumour growth (Spees et al., 2006; Tan et al., 2015). It is likely that the selective advantage provided by the new mtDNA haplotype, following mitochondrial uptake, drives its expansion (Fliss et al., 2000; Habano et al., 1998; Polyak et al., 1998).

Intercellular MT transfer has also been suggested to play a role in mesenchymal stem cell (MSC)-mediated tissue repair (Spees et al., 2016). MSC-mediated mitochondrial delivery has been observed to repair cigarette smoke and UV induced damage *in vitro* (Liu et al., 2014; Wang and Gerdes, 2015) and recent work suggests that cell damage plays a role in directing the organelle delivery process (Mahrouf-Yorgov et al., 2017). Interestingly, a previous study has shown that transfer of mtDNA *in vitro* is enhanced by the chemotherapeutic agents cytarabine, etoposide and doxorubicin (Moschoi et al., 2016).

As mentioned in the introduction (Section 1.2.1.9), phylogenetic analyses have demonstrated that multiple horizontal transfers of mitochondria have taken place in CTVT. Normally, detecting intercellular mtDNA transfer within an individual would be extremely challenging given that mtDNA haplotypes would be highly similar, if not identical. Rebbeck et al. (2011) noted at least two mtDNA horizontal transfer events in thirty seven CTVT samples originating from seven countries (Mexico, Israel, South Africa, Thailand, Kenya, Greece and Malaysia). This process has been suggested to provide a functional advantage to CTVT cells - as CTVT mtDNA acquire an increasing number of mutations with time, their efficiency may decline, and the infusion of fitter, host mitochondria, may provide either a selective advantage to CTVT cells or that competition between mitochondria leads to the selfish expansion of 'fitter' mtDNA haplotypes(Rebbeck et al., 2011). Cellular uptake of donor mitochondria could also suggest that the long-term survival of CTVT has put pressure on basic cellular metabolic processes. However, we cannot exclude that this intriguing case of mitochondrial thievery occurs, and is subsequently maintained, in CTVT by entirely neutral processes.

The exact mechanism of mtDNA horizontal transfer between cells is unknown (Berridge et al., 2015; Tan et al., 2015). Two leading theories are that mitochon-

dria transit between cells through transient cytoplasmic bridges known as tunnelling nanotubules (TNTs) or via engulfment of microvesicles containing mitochondria. Considerable evidence supports the view that TNT-like structures are involved. While shuttling of mtDNA between cells may be a normal physiological process occurring during tissue homeostasis, development and aging (Berridge et al., 2016), it is likely that horizontal mtDNA transfer occurs in response to mtDNA damage. Overexpression of Miro1, a Rho-GTPase that traffics mitochondria along an actin "highway" within tunneling nanotubes, has been shown to enhance the ability of mesenchymal stem cells to donate mitochondria and to rescue damaged epithelial cells *in vitro* as well as in a mouse model of airway inflammation (Ahmad et al., 2014).

A mechanistic explanation for how host mtDNA is captured has not been established. Given that CTVT is likely to be derived from a histiocytic or macrophage lineage (Section 1.2.1.7), it may be that tumour cells may be pre-adapted to fuse with host cells and engulf foreign material.

2.3.6 Mitochondrial recombination

Generally, mtDNA is assumed to be recombinationally inert. However, mtDNA recombination has been observed in various eukaryotes (Bergthorsson et al., 2003; Gantenbein et al., 2005; Hoarau et al., 2002; Ladoukakis and Zouros, 2001; Lunt and Hyman, 1997; Ujvari et al., 2007) as well as in human cell extracts (Thyagarajan et al., 1996) showing that mammalian mtDNA is capable of recombining. In rare cases, where paternal mitochondria has been transferred to the ooctye during fertilisation, recombination of maternal and paternal mtDNA haplotypes has been observed in humans (Kraytsberg et al., 2004; Zsurka et al., 2005).

While mtDNA recombination has previously been proposed as a mechanism for mtDNA repair (Thyagarajan et al., 1996) it has not, so far, been detected in cancer cells. One reason for this could be that mtDNA genomes in most individuals are identical given the clonal expansion of maternal mitochondria within the ovum once fertilisation occurs. It is far easier to detect occurrences of MT recombination that involve transfer of genetic material between two distinct mtDNA molecules as opposed to two similar molecules. In CTVT, multiple distinct mtDNA haplotypes may co-exist in the same cell following horizontal transfer. This situation provides a

unique opportunity to a investigate novel genetic mechanism, possibly functioning in DNA repair, previously undetected in cancer

2.3.7 Nuclear copies of mitochondrial DNA

Nuclear copies of mtDNA (NuMTs) are fragments of mtDNA that have been inserted into the nuclear genome and can be subject to subsequent genetic drift, with individual NuMTs picking up polymorphic mutations in different lineages. Rapid mobilisation of mtDNA sequences has been reported in humans (Turner et al., 2003), plants (Wang et al., 2012) and yeast (Cheng and Ivessa, 2010). Over 150 NuMTs have been identified in the canine genome (Verscheure et al., 2015) and somatically acquired NuMTs have also been described in human cancer (Ju et al., 2015). It is important in the context of this chapter to consider the possibility of mitochondrial DNA transfer into the nuclear genome. NuMT sequences are homologous to mtDNA and may be amplified alongside their mitochondrial counterparts, becoming a source of contamination in mtDNA analysis and, in particular, confound short-read analyses.

2.3.8 Summary

This introduction has given a broad overview of our current understanding of mitochondrial biology and highlighted the role of mitochondrial function in cancer and in response to cell damage. This establishes the context in which I will address the following questions related to mitochondrial function in CTVT:

(i) What is the frequency of mitochondrial capture in CTVT?

(ii) At what points in CTVT evolution have independent horizontal mtDNA transfers taken place?

(iii) What is the geographical distribution of tumours within each horizontal transfer group?

(iv) What can the inferred CTVT phylogeny along with the estimated timing of horizontal transfer events tell us about the global spread of CTVT?

(v) Can any signature of selection be detected in CTVT mitochondria?

(vi) What mutational processes affect CTVT mitochondria?

(vii) Using phylogenetic methods, can we detect any instances of mtDNA recombination in CTVT mitochondria?

Methods

The methods section is divided into two parts relating to the analysis of two data sets: (*i*) analysis of 449 complete CTVT mtDNAs (Data Set 1) generated from low-coverage whole-genome sequencing, and (*ii*) analogous analysis of 640 complete CTVT mtDNAs (Data Set 2) generated from both whole-exome and low coverage whole-genome sequencing and including 437 tumour samples from Data Set 1.

2.4 Methods (Data Set 1)

2.4.1 Sample collection and DNA extraction

Biological sample collection was approved by the Department of Veterinary Medicine, University of Cambridge, Ethics and Welfare Committee (reference number CR174). Tumour and host (gonad, skin, blood or liver) tissue samples were collected into RNAlater solution and stored at 4°C until processing. Genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue extraction kit. Sample collection was coordinated by Andrea Strakova and Elizabeth Murchison. A list of the veterinary collaborators that contributed sample material analysed in Data Set 1 is provided in the Preface. DNA extractions were performed by Andrea Strakova. Sample information is provided in Appendix 1: Supplementary file 2.1.

Sampling from matched host individuals was a vital aspect of the sample collection protocol as use of host sample sequences allowed us to identify and exclude contaminating host reads present in tumour samples. Unmatched CTVT tumours arose where collaborators were unable to provide a matched host sample or where it was not possible to successfully extract, sequence, or analyse DNA from the matched host tissue.

2.4.2 CTVT diagnosis

An initial diagnosis of CTVT was provided by veterinary collaborators based on either histopathology or clinical presentation. Diagnostic quantitative PCR (qPCR) assays were performed to confirm CTVT diagnoses via detection of a CTVT-specific LINE-MYC rearrangement (Katzir et al., 1985).

Relative levels of host contamination were estimated by qPCR amplification of a region of the cyclin-dependent kinase inhibitor 2A (*CDKN2A*) locus; since *CDKN2A* is homozygously deleted in CTVT but is present in two copies in host, amplification would be expected to correspond to host contamination. qPCR assays assessing host contamination and confirming CTVT diagnoses were performed by Andrea Strakova (Appendix 1: Supplementary file 2.3).

2.4.3 Low coverage whole genome DNA sequencing

DNA quantification and quality control were performed by the Wellcome Trust Sanger Institute Sample Management Team. Whole genome sequencing libraries were constructed using standard protocols according to manufacturer's instructions with insert size 100 to 400 base pairs (bp). Sequencing was performed on an Illumina HiSeq2000 instrument with 75 bp paired end reads. Library preparation and DNA sequencing of CTVT tumour and host samples was performed at the Core Sequencing Facility, Wellcome Trust Sanger Institute, Hinxton. Sequence reads were aligned with the CanFam3.1 dog reference genome (Lindblad-Toh et al., 2005) using the Burrows-Wheeler Aligner's (BWA) tool (v0.5.9) (Li and Durbin, 2009). Whole genome sequencing was performed to an average whole genome depth of approximately 0.3×; average mtDNA coverage was approx. 70× across 449 CTVT tumours and 338 CTVT hosts.

Samples 1380T and 1381T were extracted and sequenced separately at the Kunming Institute of Zoology, as previously described (Pang et al., 2009). DNA extractions and PCR amplifications were performed by Ting-Ting Yin according to a previously published protocol (Pang et al., 2009). The sequenced fragments were assembled using the DNASTAR SeqMan software and the complete mitochondrial genomes were aligned with the dog mitochondrial reference genome (Genbank accession U96639.2; Kim et al. 1998).

2.4.4 Mitochondrial copy number

Relative mitochondrial copy number represents the abundance of mtDNA relative to nuclear DNA. This was calculated using the following equation:

$$(COV_{MT}/COV_{nucl}) * P$$

n where COV_{MT} corresponds to average coverage across the mitochondrial genome, COV_{nucl} is average coverage across the nuclear genome and P is ploidy. CTVT tumours and hosts were treated as diploid (Murchison et al., 2014). Host and tumour samples with average MT coverage greater than 300×, which may result from degraded DNA, were excluded from the copy number calculations (Appendix 1: Supplementary file 2.2a).

2.4.5 Whole genome copy number

Simple whole genome copy number profiles were generated from low-coverage sequencing data. Briefly, for each sample we computed read depths in 10 KB windows across chromosomes 1-38 using bedtools 'coverage' v2.23.0 (Quinlan and Hall, 2010). Total copy number was inferred from read depth profiles (Fig. 2.3). Read depth plots were assessed as a quality control step in order to screen for tumour contaminated hosts and host contaminated tumours with aberrant copy number profiles. Code for this analysis was provided by David Wedge.

2.4.6 Substitution calling and filtering

Point substitutions were called using CaVEMan (Cancer Variants through Expectation Maximisation), an in-house variant calling algorithm, as previously described (Nik-Zainal et al., 2012). As CaVEMan was developed for matched tumour-normal data, and CTVT tumours and normals are unmatched (i.e. they are different individuals), I used simulated reads generated from the reference genome as the 'normal'. Single base substitutions in tumours and matched hosts were called all relative to this. Data from samples 1380T and 1381T was processed separately at the Kunming Institute of Zoology. Substitutions in 1380T and 1381T were called using MEGA (Molecular Evolutionary Genetics Analysis) (Tamura et al., 2013). Substitution calling in samples 1380T and 1381T was performed by Ting-Ting Yin.

The following list of in-built post-processing filters was applied after initial calling to improve the specificity and sensitivity of substitution calls:

- At least one third of reads supporting the mutant allele must have base quality >25.

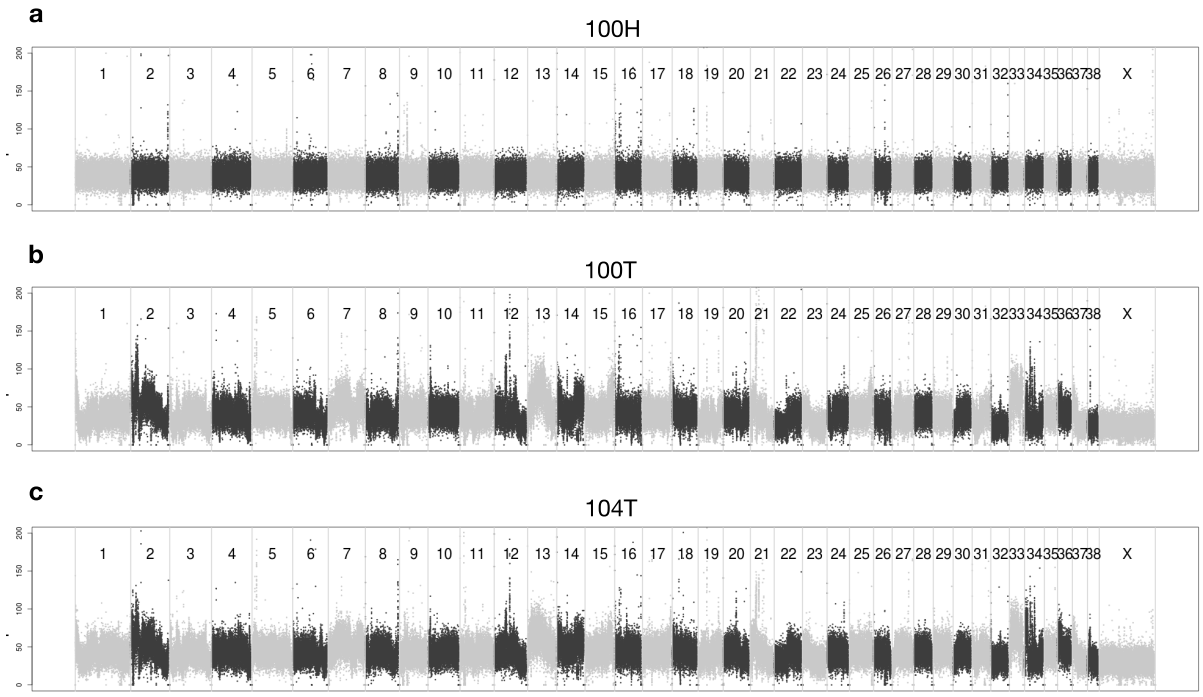


Fig. 2.3. Read depth profiles generated from low resolution whole genome sequencing used to infer total copy number for a. 100H, the matched host of tumour 100T; a characteristic host profile **b.** 100T, a clade 1 CTVT tumour sampled in Nicaragua; a characteristic tumour profile **c.** 104T, a clade 2 CTVT tumour sampled in India. Chromosomes are indicated. Read depth profile generated for 100H corresponds to normal diploid copy number.

- Mean mapping quality of reads supporting a substitution variant call must be $\geq$21.
- Substitution variant calls supported only by the first or last 15 bp of reads were discarded.
- Substitution variants were discarded if they occurred 10 bp upstream or downstream of an unfiltered indel called in the same sample (as detected by the indel-calling algorithm cgpPindel; Section 2.4.7). The 10 bp range was extended by the REP value for samples where an indel had been called with REP>0; REP represents the number of times the inserted or deleted base(s) occurs in the sequence directly 5′ or 3′ of the putative indel.
- Substitution variant calls were discarded if they occurred within the simple repeat region MT:16129-16430 inclusive. Coordinates for the dog reference genome, CanFam3.1, were obtained from UCSC table browser (http://genome.ucsc.edu/).
- Substitution variants where the wild type allele was supported by at least one read on both the forward and reverse strand but the mutant allele was supported on only one strand were discarded.

2.4.6.1 Variant allele fraction plots

A variant allele fraction was reported for each substitution detected in the CaVEMan substitution calling pipeline. VAF corresponds to the number of reads supporting the substitution variant as a fraction of the total number of reads covering the substitution variant position). Variant allele fraction plots, displaying VAF value versus mitochondrial genomic position (Fig. 2.4), were frequently used during analyses as a tool for quality control. Germline mtDNA variants detected in hosts were mostly homoplasmic and supported by 100% of reads (Fig. 2.4b). Tumour mitochondrial VAF profiles were distinct from matched host profiles, as expected (Fig. 2.4a).

2.4.6.2 Somatic substitutions in tumours with matched hosts

To remove tumour substitutions caused by host contamination, substitutions that were called in both tumour and matched host, but had VAF<0.9 in the tumour, were discarded. Substitutions with VAF>0.9 in both tumour and matched host were considered to be likely germline substitutions shared between host and tumour,

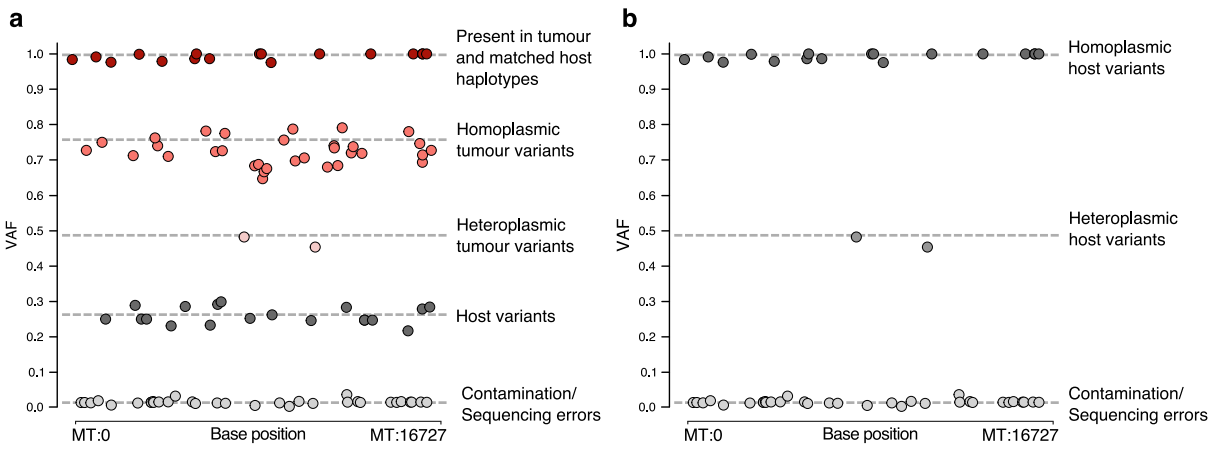


Fig. 2.4. Sample variant allele fraction (VAF) plots for a **a.** CTVT tumour and **b.** matched host. Characteristic tumour and host VAF bands are indicated. In tumour VAF plots, substitutions found in both tumour and host are present in 100% of reads. Tumour-only substitutions are present in a fraction of reads that corresponds to the amount of host contamination e.g. where host contamination is 20%, tumour-only variants are expected to be supported by 80% of reads.

and were removed. Low coverage hosts (defined as average mitochondrial genome coverage $<20\times$) were additionally checked for evidence of substitutions at positions where substitutions were called in the corresponding tumour, and the substitution was discarded in the tumour if at least one read supporting the substitution was found in the low coverage host. All tumour substitutions with VAF$<$0.5 were discarded if the matched host was of low coverage. Low-level tumour-contaminated hosts were additionally checked for the presence of substitutions identified in other tumours and any substitutions arising due to contamination were discarded.

2.4.6.3 Somatic substitutions in tumours without matched hosts

Variant allele fractions were used to identify the fraction of host contamination in each tumour as well as substitutions likely arising due to host contamination in tumours for which matched hosts were not available (Appendix 1: Supplementary file 2.1 for list of unmatched tumours). I discarded substitutions below a VAF cutoff, specified uniquely for each tumour based on its estimated level of host contamination (for most tumours, the VAF cutoff was 0.5 or 0.6). If VAF plots did not show clear distinctions between tumour substitutions and host contamination (i.e. host contamination was

greater than ~40%; this category included 9 tumours), we identified likely tumour substitutions as those which were present in phylogenetically-related tumours. Any remaining substitutions that were also found in normal dogs were removed (see Appendix 1: Supplementary file 2.4f for list of germline substitutions present in normal dogs); those substitutions that were not found either in phylogenetically-related tumours or in normal dogs were kept as putative somatic substitutions.

2.4.6.4 Germline substitutions in hosts

Filters described in Section 2.4.6 were described to substitutions in hosts. Low coverage hosts (average mtDNA coverage $<20\times$; see Appendix 1: Supplementary file 2.2a) were further checked for evidence of substitutions at positions where a substitution was called in the corresponding tumour as described in Section 2.4.6.2. Substitutions called in low coverage hosts and hosts with regions of low coverage (137H, 402H, 423H, 469H, 513H, 530H1, 764H, 82H1, 835H, 859H, 90H) may still contain false negatives due to low coverage.

2.4.6.5 Further quality checks and validation

Additional quality filtering was performed in low coverage regions, regions with low-mapping quality, and regions containing variable-length polyC homopolymer tracts (Fregel et al., 2015). Substitutions that were subsequently completely excluded from the analysis as a part of this check are as follows: 16705C>T was frequently miscalled due to low coverage in this region (the notation 16705C>T is used to indicate a G>A substitution at MT base position 16705); 15931A>G, 16672C>T were frequently miscalled in samples due to their presence in the middle of a homopolymer tract; 15632C>T, 15639T>A, 15639T>G were frequently miscalled due to poor mapping quality in this region;15493G>A, 15505T>C were inconsistently called across samples due to low coverage and poor mapping quality in this region.

Substitutions that were discarded due to proximity to an indel (Section 2.4.6) were visually inspected using Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdóttir et al., 2013). The following subset of these substitutions had substantial support and were rescued: 381T>A, 1481T>C, 1683T>C, 2682G>A, 2683G>A, 3028A>C, 6629T>C, 6882A>G, 7014T>G, 8281T>C, 8368C>T, 8703G>A, 9825G>A, 9896T>C, 13708C>T, 14977T>C, 15524C>T, 15526C>T, 16660T>C, 16663C>T, 16671T>C.

2.4.6.6 Host contamination

Host contamination levels in each tumour were estimated from VAF plots (Section 2.4.6.3). Substitutions that were present in tumours but not in matched host were identified, and their average VAF used to estimate the proportion of tumour mtDNA. Substitution VAFs were normalised to take account of host contamination. While it is most likely that tumour variants with normalised VAF< 1 represent heteroplasmic variants, we cannot exclude that these represent cellular subclones harbouring distinct homoplasmic mtDNA populations.

2.4.6.7 Recurrent mutations and back mutations

Back mutations are two mutations (either two somatic mutations or a germline and a somatic mutation) at the same position leading to a reversion to the reference allele. Recurrent mutations are when the same somatic variant arises independently in different CTVT tumours or when a somatic variant occurs at the same position as a germline variant.

Back mutations and recurrent mutations occurring in tumours were identified by inspecting positions of tumours carrying each substitution on phylogenetic trees (See Appendix 1: Supplementary file 2.15 for a summary of back mutations).

2.4.6.8 Extracting substitution variants from publicly available dog sequences

In order to enrich the germline substitution panel created from 338 CTVT hosts, an additional 252 publicly available complete dog mtDNA genomes were included. Freely available complete dog mtDNA sequences were downloaded from the National Center for Biotechnology Information (NCBI). To extract substitution variants, FASTA sequences were aligned with the dog mitochondrial reference genome (Genbank accession U96639.2) using Clustal Omega (Sievers et al., 2011). SNP-sites (Page et al., 2016) was used to call variants in the multiple sequence alignment. Alignment errors, usually due to miscalls caused by closely mapped indels, were inspected manually and corrected to minimise gaps. For those samples missing data in regions MT:15510-15532 and MT:16040-16550 I substituted the most likely substitution at polymorphic sites based on phylogenetic position. The previously outlined filtering

rules (Sections 2.4.6) were applied to the substitution set where applicable. Substitutions called before MT:48 or after MT:16671 were excluded due to low coverage in these regions of sequencing data. Substitutions represented by International Union of Pure and Applied Chemistry (IUPAC) codes R, Y, S, W, K or M where one of the two possible calls was the same as the reference were changed to the base which was different to the reference. In cases where the IUPAC code represented >2 bases (B, D, H, V, or N), the reference base was substituted.

2.4.7 Indel calling and filtering

Small insertions and deletions (indels) were extracted from the sequencing data using cgpPindel, a modified version of Pindel (Ye et al., 2009). The following list of in-built filters was used to improve the specificity and sensitivity of indel calls:

- Indels were required to have ≥3 supporting reads on either the forward or reverse strands or ≥2 supporting reads on both the forward and reverse strands
- Indel calls with at least 4 supporting pindel-mapped reads were required to have at least 1 supporting BWA-mapped read or, failing that, if REP= 0 (Section 2.4.6), then at least one supporting pindel-mapped read on both strands.
- Indels within the simple repeat region MT:16129-16430 inclusive were discarded. Coordinates for the dog reference genome, CanFam3.1, were obtained from UCSC table browser (http://genome.ucsc.edu/).

Samples with very high coverage of the mitochondrial genome (1T, 2T, 3T, 4T, 4H, 498H, 432T, 455T1, 231T, 79H, 79T) were excluded from this analysis due to an increased false discovery rate.

Samples 1380T and 1381T were processed separately at Kunming Institute of Zoology. Indels were not called in these samples and they were not included in this analysis.

2.4.7.1 Somatic indels

Indels called in tumours that were also called in at least one host were discarded as possibly arising due to host contamination. Indels that were uniquely called in tumours without matched hosts were discarded, as the possibility that they were caused by host contamination could not be ruled out. All remaining indels were

Table 2.1 Indels discarded from CTVT mitochondrial variant analysis

Base Position	Indel	Sample	Basis for discarding indel
9891	TGATTTATCTCATAATTATCA>TATCTCATAATTATCATG	324T2	False positive called due to close proximity to two other indels
9910	C>CAT	401T	False positive called due to the presence of a substitution at the same position

visually validated using IGV (Robinson et al., 2011; Thorvaldsdóttir et al., 2013). The indels listed Table 2.1 were discarded from the analysis as miscalls. In total, 27 somatic indels were included in the analysis, including recurrent indels (Appendix 1: Supplementary file 2.5a).

2.4.7.2 Germline indels

Germline indels were those which were considered homoplasmic (average VAF$\geq$0.9) in CTVT hosts and which passed a visual validation performed using IGV (Robinson et al., 2011; Thorvaldsdóttir et al., 2013). In total, 7 CTVT host homoplasmic germline indels were identified (Appendix 1: Supplementary files 2.5a and 11).

2.4.7.3 Recurrent indels

Recurrent indels occurring in tumours were identified by inspecting phylogenetic positions of tumours carrying each indel.

2.4.7.4 Calculating Indel Allele Fraction

Although primary indel calling was carried out using cgpPindel, indel allele fraction for wild type and mutant indels was calculated using vcfCommons (in-house software developed at the Wellcome Trust Sanger Institute). The algorithm takes all mapped and unmapped reads in the region of an indel and performs an alignment of the reference and the predicted mutated path using Exonerate (Slater and Birney, 2005). The predicted mutated path is the reference with the change predicted by cgpPindel applied to it. Based on this alignment, reads are classified into 3 categories:

(i) aligns to reference path

(ii) aligns to mutated path

(iii) ambiguous; a read sequence aligns to the reference and mutated path with identical scores which makes it impossible to determine the true path

2.4.8 Phylogenetic analyses

Phylogenetic analyses were used to infer evolutionary relationships between CTVT tumours. Trees were constructed using a maximum likelihood (ML) method implemented in PhyML 3.0 (Guindon et al., 2010) using the General Time Reversible (GTR) nucleotide substitution model allowing for the transition/transversion (ts/tv) ratio, gamma distribution shape parameter and proportion of invariant sites to be estimated in PhyML for each data set. Phylogenetic trees constructed using RAxML v8.2.6 with a rapid hill-climbing algorithm and identical model specification were consistent with the phylogenetic trees obtained using PhyML. Prior to tree construction, jModelTest2 (Darriba et al., 2012) was used to determine the best fitting substitution model for the alignment, inferred using the Akaike Information Criteria (AIC). The tree topology was estimated using a combination of Nearest Neighbor Interchange (NNI) and Subtree Pruning and Regrafting (SPR) (Hordijk and Gascuel, 2005) algorithms. Trees were visualised using Dendroscope (v3.2.10; Huson and Scornavacca, 2012). Bootstrap support values were obtained from 100 bootstrap replicates. The long branch-lengths observed in CTVT clade 4 were further investigated by visually validating sequence alignments involving these samples. Source data is available in Appendix 1: Supplementary files 2.12, 2.13 and 2.14.

2.4.9 Confirming independent mtDNA horizontal transfer events

2.4.9.1 Classification of tumour substitutions

CTVT tumour substitutions were classified based on their inferred origin as follows:

Germline clade-defining substitutions
Tumour germline clade-defining substitutions are those identified within the pool of germline substitutions (n = 1152; Appendix 1: Supplementary file 2.4f) in the sampled normal dog population (338 CTVT host samples and 252 publicly available dog mitochondrial genomes). There variants were inferred to be present on the donor dog mtDNA haplotype which founded an individual horizontal transfer event and, by definition, these variants are shared by all tumours within a phylogenetic

grouping or clade derived from the original CTVT cell which received the horizontally transferred donated mtDNA. A complete list of tumour germline clade-defining substitutions is provided in Table 2.2.

Somatic substitutions
Tumour somatic substitutions are variable within the set of tumours in one clade and thus are phylogenetically informative. These variants are not present in the current dog population and were inferred to have arisen somatically after the horizontal transfer event that founded a clade (see Appendix 1: Supplementary file 2.4b for a complete list). Substitutions associated with the inferred recombination event in 559T (i.e. present uniquely in 559T) were excluded (Section 2.4.12).

Conservative somatic substitutions
Conservative tumour somatic substitutions were variable and phylogenetically informative within the set of tumours in one clade and inferred to have arisen after the clade-defining horizontal transfer event (as above). However, somatic substitutions within this set which had the potential to be germline were excluded (see Appendix 1: Supplementary file 2.4c for a full list of 'conservative' somatic mutations). In order to curate a set of high-confidence or 'conservative' somatic substitutions the following variant sets were excluded:

- Variants associated with inferred recombination events (Section 2.4.12).
- Variants uniquely present in clade 3; clade 3 tumours carry very few somatic mutations, and so the possibility that clade 3 tumours arose from several independent mtDNA horizontal transfer events cannot be excluded
- Variants occurring on basal trunks of the CTVT mtDNA tree; these include trunks leading to haplogroups CTVT_1A, CTVT_1B1 and CTVT_2A (labelled in Fig. 2.7). Due to their ancestral divergence, it cannot be excluded that these haplotypes are derived from independent mtDNA horizontal transfer events
- Putative somatic substitutions variants in tumours without hosts (Section 2.4.6.3).

This set of substitutions was used for the analyses described in Section 2.4.13 and Section 2.4.14.

Potential somatic substitutions
Tumour potential somatic substitutions are those common to all tumours within a

single clade but not present in the pool of 590 normal dogs analysed (Appendix 1: Supplementary file 2.4f). It could not be determined if these substitutions were rare germline variants present on the 'donated' dog mtDNA genome that founded each clade, or if these were early somatic substitutions that arose after the clade-defining horizontal transfer event, but prior to the divergence of tumours analysed in this chapter. See Table 2.2 for a full list of tumour potential somatic substitutions for all clades identified in Data Sets 1 and 2.

2.4.9.2 Characterising CTVT clades

A CTVT clade was defined as a group of tumours deriving from a single horizontal transfer event. Independent clades were defined on the basis of the tumour substitution classification outlined above as follows:

1. CTVT tumours from the same clade arose from a single donor mtDNA haplotype and therefore share the same set of germline substitutions which was inferred to be originally present on the donor mtDNA haplotype (germline clade-defining substitutions; Section 2.4.9.1).
2. CTVT tumours from the same clade clustered together on the phylogenetic tree (Fig. 2.6).
3. The reconstructed donor mtDNA haplotype for each clade has a phylogenetically closely related or identical mtDNA haplotype in the sampled dog population.

2.4.10 Haplotype nomenclature

2.4.10.1 Host

The host haplotype naming system used in this chapter was adapted from the cladistic canine mitochondrial DNA phylogeny nomenclature proposed by Fregel et al. (2015). See Appendix 1: Supplementary file 2.9 for host haplotype names. Corresponding CTVT normal dog host samples were assigned into one of the major clades (A, B, C, D, E and F; Section 2.3.2.1). For subsequent levels, haplogroups were defined by specific diagnostic variants, as defined by Fregel et al. (2015). A unique number, following an underscore after the haplogroup name, distinguished distinct haplotypes within each haplogroup. Haplogroup-defining variants 15632C>T, 15639T>A and 15639T>G

Table 2.2 CTVT germline and potential somatic clade-defining substitutions (Data Set 1).
Germline clade-defining substitutions are those shared between all samples within a clade and also present in the pool of dog host substitutions. There are no potential somatic clade-defining substitutions for clades 2 and 3.

CTVT Clade 1	CTVT Clade 2	CTVT Clade 3	CTVT Clade 4	CTVT Clade 5
2683 G>A	5367 C>T	1351 A>G	295 G>C	229 C>T
2833 C>T	5444 T>C	2683 G>A	395 A>T	2322 T>A
2962 C>T	6065 A>G	2962 C>T	530 T>C	2683 G>A
3196 T>C	6257 G>A	3196 T>C	885 G>A	2962 C>T
3641 G>A	8368 C>T	4940 T>C	955 C>T	3196 T>C
4591 G>A	8807 G>A	5367 C>T	1481 T>C	5367 C>T
5367 C>T	13299 T>A	5444 T>C	2683 G>A	5444 T>C
5444 T>C	15814 C>T	6065 A>G	3628 A>T	6065 A>G
5700 G>A	16025 T>C	6401 C>T	3897 A>T	6401 C>T
6065 A>G		6554 T>C	3936 T>C	6494 A>G
6302 A>G		7593 T>C	5367 C>T	7324 G>A
8242 G>A		8281 T>C	5444 T>C	7593 T>C
8281 T>C		8368 C>T	6065 A>G	8281 T>C
8368 C>T		8807 G>A	6068 T>C	8368 C>T
8807 G>A		10611 A>T	6744 A>T	8408 C>T
9790 G>A		10992 G>A	8368 C>T	8807 G>A
9896 T>C		13299 T>A	8565 T>C	8961 A>G
10304 G>A		14977 T>C	8807 G>A	9372 C>T
10992 G>A		15214 G>A	10313 G>A	10467 A>T
11657 C>A		15620 T>C	10618 A>T	10497 A>T
12505 T>C			11945 C>T	10992 G>A
13299 T>A			12330 A>G	11975 G>A
15214 G>A			12336 G>A	12345 C>T
15627 A>G			13299 T>A	12290 C>T
15814 C>T			13620 C>T	12858 G>A
16025 T>C			13971 A>T	13299 T>A
			15814 C>T	14554 A>G
			16025 T>C	14977 T>C
			15214 G>A	
			15613 A>G	
			15627 A>G	
			15814 C>T	

were excluded from this analysis (Section 2.4.6.5) and, thus it was not possible to distinguish between haplogroups A1c and A1e. Haplotypes which did not fit into any haplogroup were classified as 'unassigned'. Haplotypes of hosts with low coverage ($<20\times$) and low coverage regions were individually assessed and assigned to haplogroups.

2.4.10.2 Tumour

We devised a CTVT mtDNA haplotype naming system adapted from the cladistic canine mtDNA phylogeny nomenclature proposed by Fregel et al. (2015). See Appendix 1: Supplementary file 2.9 for a complete list of tumour haplotype names. Each CTVT haplotype has a prefix 'CTVT_', indicating a tumour haplotype. The five CTVT clades are numbered (CTVT_1, CTVT_2, CTVT_3, CTVT_4 and CTVT_5). For subgroups within clades, hierarchical notation was used, where subsequent haplogroups are distinguished by alternating letters and numbers and the maximum number of levels included in the hierarchical notation was five – i.e. 3 numbers and 2 letters (e.g. 1A1a1). The first letter is an upper case Roman letter; subsequently-used letters are lower case Roman letters. Any haplogroups beyond the maximum number of levels were considered as a single subgroup, in which individual haplotypes were distinguished using a non-hierarchical numbering system - an underscore followed by a number (e.g. 1A1a1_1, 1A1a1_2, etc.). Underscores were only used to distinguish individual haplotypes after the haplotype has been assigned to all 5 hierarchical levels (e.g. 1A_1 does not exist, as this haplotype would be classified as 1A1 instead).

2.4.10.3 Reconstructed donor haplotypes

A 'donor haplotype' was reconstructed for each of the clades, representing the inferred donor mtDNA haplotype in each horizontal transfer event, and was used to root the trees for each clade. Donor haplotypes were reconstructed from the clade-defining germline substitutions and the clade-defining potential somatic substitutions (Section 2.4.9.1). The phylogenetically closest haplotype present in the current dog population is shown in the same figure. The clade-defining germline variants represent substitutions present on the reconstructed donor mtDNA.

2.4.11 Estimated timing of clade divergence

The timing of clade divergences was estimated using the following three independent methods:

- timing based on nuclear DNA (Alexandrov et al., 2013; Murchison et al., 2014),
- timing based on number of cell divisions per homoplasmic mitochondrial mutation (Cohen and Steel, 1972; Ju et al., 2014), and
- timing based on number of mitochondrial mutations per year (Ju et al., 2014).

All estimated divergence times assume a constant rate of accumulation of somatic mtDNA mutations both within and between clades as well as constant activity of selection. The average number of somatic mutations within each clade was calculated for each of the five clades (Tables 2.4 and 2.5). As the number of somatic mutations influences the estimated time of divergence, estimates both with and without potential somatic mutations are included in this analysis (Tables 2.4 and 2.5).

The presence of mitochondrial recombination in haplogroup CTVT_1B2b2 (CTVT clade 1) was taken into account when calculating the average number of mutations per clade (Section 2.6.8).

2.4.11.1 Timing based on nuclear DNA

Based on the most recent common ancestor of samples 24T (identified in the previous analysis as a CTVT clade 1 sample) and 79T (identified in the previous analysis as a CTVT clade 2 sample) which is estimated to have existed approximately 460 years ago (Murchison et al., 2014), it can be assumed that the maximum age of CTVT clade 2 (the more recent of CTVT clades 1 and 2) is 460 years. Calibrated according to this dating (Alexandrov et al., 2013; Murchison et al., 2014) and assuming a constant rate of accumulation of mutations with time, the maximum number of CTVT somatic mtDNA mutations per year is 0.0205 (calculated using average number of mutations in clade 2 = 9.437). The calculated maximum time since origin of clades 1, 3, 4 and 5 is shown in Tables 2.4 and 2.5.

2.4.11.2 Timing based on number of cell divisions

A study of mtDNA mutations in human cancers estimated that one homoplasmic somatic mtDNA mutation arises every ~1000 cell generations in human cancers (Ju

et al., 2014). An experimental study estimated CTVT cell generation times to be 4 days for first stage tumours and >20 days for second stage tumours (Cohen and Steel, 1972). Using 4 days and 20 days as a minimum and maximum generation time, and assuming a constant somatic mtDNA mutation rate in CTVT, we estimated a minimum and maximum mutation rate of ~0.0183 and ~0.0913 mutations per year respectively. See Tables 2.4 and 2.5 for corresponding calculations of minimum and maximum time since clade origins.

2.4.11.3 Timing based on number of mtDNA mutations per year

A previous study correlated somatic mtDNA mutation accumulation in human cancers with patient age (Ju et al., 2014). This suggested an approximate rate of ~0.025 mutations per year. Assuming a similar rate in CTVT somatic mtDNA mutations and a constant CTVT somatic mtDNA mutation rate, the time since mtDNA horizontal transfer events was estimated (Tables 2.4 and 2.5).

2.4.12 Mitochondrial recombination analyses

We screened for potential mtDNA recombination in CTVT tumours using phylogenetic methods. If recombination had occurred in CTVT mtDNA then the expectation would be that phylogenetic trees generated using different regions of the mitochondrial genome would result in incongruent phylogenies.

2.4.12.1 Automated recombination analysis

The RDP4 package (Martin et al., 2015) was used to detect recombination events within the complete sample set (449 CTVT tumours, 338 CTVT hosts and 252 additional dogs) using a Bonferroni corrected p-value cutoff of 0.05. Default parameters were used for the following programs implemented within the RDP package: RDP (Martin and Rybicki, 2000), MaxChi (Maynard Smith, 1992), Chimaera (Posada and Crandall, 2001), 3Seq (Boni et al., 2007) and SiScan (Gibbs et al., 2000).

2.4.12.2 PacBio long read sequencing

The PacBio single molecule, real-time (SMRT) sequencing technique was used to validate the finding of recombinant mtDNA sequences. SMRT sequencing can generate ultra-long reads (>30,000 bp) with extremely high consensus accuracy and this

technique had previously been used to sequence whole mitochondrial genomes (Ma and O'Farrell, 2015).

A genomic library was created directly using $5\,\mu g$ of genomic DNA from sample 559T, not utilising shearing or amplification techniques, as previously described (Coupland et al., 2012). The library was sequenced using two PacBio SMRT cells using the Pacific Biosciences RS sequencer. Each SMRT cell yielded ~1 Gb of sequence data with mean read length 11,421 bp and N50 read length 19,382 bp. Average sequence coverage across the mitochondrial genome was $111.3\times$. PacBio long read sequencing was overseen by Andrea Strakova. Library preparation and sequencing was performed at the Core Sequencing Facility, Wellcome Trust Sanger Institute, Hinxton.

2.4.12.3 PacBio data analysis

PacBio sequence reads aligning to mtDNA were viewed in SMRT view as well as in Integrative Genomics Viewer (IGV) (Robinson et al., 2011; Thorvaldsdóttir et al., 2013) and used to phase the mitochondrial substitutions previously called in 559T. This analysis was performed by Andrea Strakova. The three most common recombinant haplotypes were completely phased (Fig. 2.12). Additional haplotypes, which were present at very low levels (less than 5%), were also identified. However, it was not possible to phase these completely. Reads derived from 559H, the host of 559T (haplotype B1_1), were also identified.

2.4.13 Selection analyses

Next, we assessed the functional effects of somatic mtDNA mutations and searched for evidence of selection by comparing variant allele fraction across different categories of somatic mutations and using dN/dS estimates. The details of these analyses are outlined below.

2.4.13.1 Substitution and indel variant allele frequency

Variant effect predictor (VEP) (McLaren et al., 2010) was used to predict the functional impact of single point substitutions and indels (see Appendix 1: Supplementary files 2.6 and 2.7). Normalised VAF values for somatic substitutions and indels were calculated as described in Section 2.4.6.6. Cumulative distribution functions of normalised

variant allele fractions were plotted for nonsense (n = 10), missense and synonymous substitutions (n = 610), using the conservative somatic list (Section 2.4.9.1; Appendix 1: Supplementary file 2.4c). Cumulative distribution functions of normalised variant allele fractions were plotted for frameshift (n = 18) and non-frameshift (n = 9) indels. Statistical significance was tested using the two-sample Kolmogorov-Smirnov test implemented in R (Team, 2013).

2.4.13.2 dN/dS

A widely used measure of selection is the dN/dS ratio; the ratio of non-synonymous (dN) to synonymous substitutions (dS). dN/dS was estimated using a method adapted from Martincorena et al. (2015). Briefly, a context-dependent model was used with 192 substitution rates (12 possible substitution types C>A, C>G, C>T, T>A, T>C, T>G, A>C, A>G, A>T, G>A, G>C, G>T x 4 possible 5′ bases x 4 possible 3′ bases), accounting for any confounding context-dependent effects by 1 nucleotide upstream and downstream. The substitution rate was modelled as a Poisson process where the product of the underlying mutation rate and the impact of selection give the rate. A likelihood ratio test was used to test the deviation from neutrality (w_{MIS}=1 or w_{NON}=1), giving a p-value for the evidence of selection. To avoid any confounding effects due to the highly strand-biased CTVT mtDNA mutation spectrum (Section 2.4.14), *ND6*, the only mitochondrial-encoded gene transcribed from the light strand (Section 2.3.2), was excluded from this analysis.

Each mutation type is modelled as a Poisson distributed random variable with a rate given by the product of the mutation rate and the impact of selection. By calculating the mutability of each trinucleotide in the coding strand, the model also accounts for immediate context-dependent mutagenesis and transcriptional strand bias in mutation rates.

2.4.14 Mutation spectrum

As described in the introduction to this chapter, mtDNA is composed of a heavy and light strand. The reference strand in CanFam3.1 corresponds to the light strand. Each mutation on the conservative somatic list (n = 835; Appendix 1: Supplementary file 2.4c) was classified as one of six possible substitutions in the pyrimidine context (C>A, C>G, C>T, T>A, T>C, T>G) and assigned to a strand relative to the reference

(i.e. pyrimidine mutations (i.e. C>, T>) with respect to the reference were defined as light strand mutations; purine mutations (i.e. A>, G>) with respect to the reference were defined as heavy strand mutations). The immediate 5′ and 3′ sequence contexts for each CTVT mutation was extracted from the dog mitochondrial reference genome for mutations on the heavy and light strands, yielding a maximum of 96 mutation types (6 possible substitutions x 4 possible 5′ bases x 4 possible 3′ bases).The number of observations of each substitution type was normalised to the triplet frequency extracted from the canine mitochondrial genome.

The following example illustrates how to calculate the observed/expected ratio for T[C>T]G occurring on the heavy strand. A total of 835 substitutions occurring across the MT genome was observed; given that the TCG triplet is observed 117 times in the dog mitochondrial reference genome heavy strand, the frequency of TCG triplets occurring in the dog mtDNA heavy strand is 117/16727 = 0.007, where 16,727 bp is the length of the dog mitochondrial genome. Using the frequency of TCG triplets in the reference genome, the expected number of T[C>T]G substitution types on the heavy strand can be calculated as (total number of mutations) x (TCG frequency on the heavy strand) / 3 (as there are 3 possible C>N substitutions) i.e. expected number of T[C>T]G substitutions on the heavy strand = 835 x 0.007/ 3 $\approx$ 1.95. As we observed 22 T[C>T]G mutations on the heavy strand, the observed/expected ratio for this mutation type was 22/1.95 = 11.28.

Triplets within region MT:16129–16430 inclusive (Section 2.4.6), as well as a set of specific sites (Section 2.4.6.5) were excluded from this analysis. This was accounted for during the calculation of expected substitutions described above.

2.4.15 NuMTs

Finally, in this subsection I describe analysis carried out to evaluate the contribution of NuMTs (Section 2.3.7) to the CTVT mtDNA variant catalogue.

To assess the potential contribution of NuMTs present within the CanFam3.1 assembly to variant calling, wgsim (https://github.com/lh3/wgsim) was used to simulate sequence reads from CanFam3.1 (excluding the MT chromosome) to a coverage of 0.3$\times$ (i.e. the average nuclear genome sequence coverage obtained for

Data Set 1). Reads were then aligned to the MT reference sequence using BWA (Li and Durbin, 2009) and Samtools depth (Li, 2011) was used to assess MT genome coverage. Any MT genome coverage detected from this analysis would be expected to arise from NuMTs.

2.5 Methods (Data Set 2)

In the following subsections I describe implementation of an improved variant calling pipeline developed using the Platypus tool. A new pipeline was designed, largely in order to minimise the amount of manual curation of mtDNA variants required (Section 2.4.6.5), and applied to an extended data set of mtDNA sequences from whole-genome and whole-exome sequencing data.

2.5.1 Summary of Data Set 2

Data Set 2 comprised of 640 CTVT tumours and 494 CTVT hosts from 44 countries collected between 2001 and 2016. This sample set includes 437 of the CTVT tumours analysed previously (whole exome and whole genome data was available for 338 of these samples) and an additional 203 samples for which only whole exome data was available. 304 hosts from Data Set 1 are also included in this analysis.

Sample collection and sequencing data related to Data Set 2 will become publicly available upon publication of Baez-Ortega *et al*. (2018), currently under review (see Appendix 2).

2.5.2 Sample collection and DNA extraction

Samples were collected from gonad, skin, blood or liver tissue into RNAlater and stored at 4°C until processing. Sample collection was coordinated by Andrea Strakova and Elizabeth Murchison. DNA extractions were performed by Andrea Strakova as described in Section 2.4.1. A list of the veterinary collaborators that contributed sample material analysed as part of Data Set 2 is provided in the Preface.

2.5.3 Confirmation of CTVT diagnosis

CTVT diagnoses were confirmed using qPCR assays as previously described (Section 2.4.2).

2.5.4 Exome Pull-Down and Sequencing

Hybrid-capture target enrichment was performed using a custom Agilent exome capture kit. The canine whole exome enrichment kit was designed based on the latest CanFam3.1 reference genome. The total size of the design was 43 Mb across 190,958 regions.

Whole genome sequencing libraries were constructed with insert size 100 to 400 bp, as previously described (Section 2.4.3). Captured DNA was sequenced on an Illumina HiSeq 2000 instrument with 75 bp paired end reads. Average depth coverage ranged from 30× to 100×. Reads were aligned against the CanFam3.1 canine reference genome using BWA (Li and Durbin, 2009).

Although this exome capture kit was not designed to target canine mtDNA genes, the high copy number of mtDNA in mammalian cells (Section 2.3.1) makes it possible to obtain mtDNA sequencing data using off-target reads coming from the MT genome (Ju et al., 2014).

Phylogenetic analysis of the CTVT nuclear genome, using the exome sequencing data included in Data Set 2, was carried out by Adrian Baez-Ortega.

2.5.5 Variant calling and filtering

Substitutions and short indels were called using Platypus v0.8.1 (Rimmer et al., 2014) (https://github.com/andyrimmer/Platypus), a haplotype-based variant caller, implemented as part of a custom computational pipeline (Mitotypus) tailored to call somatic and germline variants in unmatched tumour and host tissues. Platypus produces genotype calls, and so offers a huge advantage over CaVEMan and cgpPindel, the variant calling softwares used in the analysis of Data Set 1 (Section 2.4.6). This population-based variant calling approach ensured that, by sharing information across samples during the genotyping step, variants with low read support in a

subset of samples, or even a single sample, could be called based on robust support for that variant in another sample. This strategy enhanced variant calling capabilities in regions of low coverage. No additional post-processing steps were required for low coverage samples unlike the CaVEMan variant calling pipeline, as described in Section 2.4.6.5. In this pipeline, the genotyping step allowed us to call any variants present at low-allelic frequency in low coverage samples, which would have appeared as false negatives before.

The design rationale of the initial variant calling and filtering steps outlined below is based largely on Somatypus (Somatypus v1.3; https://github.com/baezortega/somatypus), also a Platypus based variant caller developed by Adrian Baez-Ortega.

Individual variant calling
Variant calling was carried out in two stages on a sample-by-sample basis. In the first execution, default settings were used along with the following requirements

- $\geq$3 supporting reads with mapping quality $\geq$ 20 were necessary for a substitution or indel call to be called (–minReads=3).
  For SNPs, the base qualities of supporting reads must be $\geq$ 20.
- Variant calls were not excluded on the basis of their Phred score value (–minPosterior=0)

The second execution was run with default settings and with the following options

- –minReads=3 (see above)
- –minPosterior=0 (see above)
- Indel candidates were not excluded on the basis of the number of reference-matching bases on either side of the indel site (minFlank=0)
- Base-qualities within 10 bp of the end of reads were set to 0 (–trimReadFlank=10)

Filtering variant calls
In-built filters, listed below were used to filter variants prior to genotyping. These filters specifically targeted variants with either low base quality, low mapping quality, low Phred quality scores, a skewed strand distribution or within low sequence complexity regions.

- *Bad Reads*: Variants were discarded if supporting reads within 7 bp upstream or downstream of the variant call site had a median base quality <15 .

- *Mapping Quality*: Variants were excluded where the root-mean-square mapping quality across all reads mapped to a variant site was below 40. The mapping quality statistic was computed using all mapped reads, before any filtering.
- *Strand Bias*: Reads supporting a variant call were compared to a binomial distribution where the mean of the distribution equals the ratio of forward and reverse strands observed in all reads. Variants were rejected if the p-value<0.001.
- *Sequence Complexity*: A sequence complexity statistic was computed measuring the contribution of the two most frequent nucleotides among the 21 flanking a variant site; if this measure exceeded 95%, variant calls were flagged and discarded.
- *Quality over depth*: Variant calls where the ratio of variant quality (Phred score) to number of supporting reads was below 10 were excluded.

Filtered SNV and indel calls from all samples then were merged. Only bi-allelic indels were included in the filtered indel call set.

Genotyping
To allow for multi-allelic SNVs (with up to three observed alleles differing from the reference allele) SNV genotyping was split into three steps. Substitution variants called ≤ 5 bp upstream or downstream of an indel (SNVs-near-indels) in any sample were extracted from the SNV data set and genotyped separately. Filtered and merged indels were genotyped across the entire sample set. Indel genotyping involved a single step.

Generating a final variant set
Variants flagged during genotyping by any of the quality checks outlined in '*Filtering variant calls*' were discarded. Any indel flagged during the genotyping step was excluded from all samples. SNVs-near-indels with median read coverage <20$\times$ or median VAF <0.2 or >0.9 were discarded.

2.5.5.1 Somatic substitutions in tumours with matched hosts

Substitutions called in both tumours and matched hosts with VAF<0.9 were discarded as a means of filtering tumour substitutions likely arising from host contamination. Substitutions with VAF>0.9 in tumour and matched host were retained.

2.5.5.2 Somatic substitutions in tumours without matched hosts

Tumours matched to host samples contaminated with tumour tissue were treated as tumours without matched hosts and the corresponding host was discarded from the analysis. Tumour samples with evidence of high host contamination were also discarded from the analysis. Variant allele fraction versus MT genome coordinates were plotted for each sample and used these to determine the level of contamination (tumour or otherwise) in each host sample as well as the level of contamination (host or otherwise) in each tumour sample. For example, in a pure host sample we would expect to see a band of homozygous SNPs at VAF = 1.0. In the case of tumour contaminated host samples, this band will instead appear at VAF<1.0. The level of tumour contamination corresponds to the gap between the expected VAF band at 1.0 and the observed VAF band at less than 1.0. By looking at the median VAF of all variants called in a host sample after filtering, we were readily able to detect sample contamination.

2.5.6 Phylogenetic analysis

2.5.7 Additional filtering, quality checks and validation

The control region of the canine mitochondrial genome is a non-coding region approximately 1200 bp in length formed of two hypervariable regions flanking a simple repeats region. In some literature this is referred to as the Variable Number of Tandem Repeats (VNTR) region. This region consists of a 10 bp tandem repeat 5'-GTACACGT(A/G)C-3' with variable repeat numbers, both between and within individuals (Kim et al., 1998).

Variants called in the simple repeats region (MT:16129-16430) of the MT genome were discarded. Variants called in the region MT:15524-15535 were also discarded. I identified the expansion of a poly-C tract, 5'-TCCCCTCCCCT-3', in this region resulting in alignment problems and leading to false positive variants. Two further homopolymeric regions were identified between MT:15926-15938 (5'-TTTTTAGGGGGGGAA-3') and MT:16662-16672 (5'-CCTTTTTTTTCC-3') which were excluded from the analysis. I also excluded all variant calls in the region MT:1-75, i.e. one read-length from the beginning of the MT contig, where we identified the occurrence of false negatives

Table 2.3 CTVT germline and potential somatic clade-defining substitutions (Data Set 2).
These substitutions are shared between all samples within a clade and also present in the pool of substitutions found in normal dogs(Section 2.4.9.1). Potential somatic clade-defining substitutions appear in italics. These substitutions are shared between all samples within a clade but absent from the pool of substitutions found in normal dogs. * indicates clade-defining variants that are absent from recombinant clade 1 samples. † indicates clade-defining variants that are absent from recombinant clade 3 samples (Section 2.6.8).

CTVT Clade 1	CTVT Clade 2	CTVT Clade 3	CTVT Clade 4	CTVT Clade 6	CTVT Clade 7
*2683 G>A	5367 C>T	1351 A>G	*295 G>C*	*615 G>A*	2683 G>A
*2833 C>T	5444 T>C	2683 G>A	*395 A>T*	*807 T>C*	5367 C>T
*3196 T>C	6065 A>G	2962 C>T	*530 T>C*	*1084 A>G*	5444 T>C
**3641 G>A*	6257 G>A	3196 T>C	*885 G>A*	*1405 G>A*	6065 A>G
*4591 G>A	8368 C>T	4940 T>C	*955 C>T*	*2036 T>C*	8368 C>T
5367 C>T	8807 G>A	5367 C>T	*1481 T>C*	2683 G>A	8536 C>T
5444 T>C	13299 T>A	5444 T>C	2683 G>A	2833 C>T	10165 C>T
5700 G>A	15814 C>T	6065 A>G	*3628 A>T*	2962 C>T	13299 T>A
6065 A>G	16025 T>C	6401 C>T	*3897 A>T*	3196 T>C	14474 G>A
6302 A>G		6554 T>C	*3936 T>C*	*3641 G>A*	15639 T>A
8242 G>A		7593 T>C	5367 C>T	*3934 T>C*	15814 C>T
8281 T>C		8281 T>C	5444 T>C	4591 G>A	16025 T>C
8368 C>T		8368 C>T	6065 A>G	5105 C>T	
8807 G>A		8807 G>A	6068 T>C	5367 C>T	
9790 G>A		10611 A>T	*6744 A>T*	5444 T>C	
9896 T>C		10992 G>A	8368 C>T	*5700 G>A*	
10304 G>A		13299 T>A	*8565 T>C*	5712 G>A	
10992 G>A		14977 T>C	8807 G>A	6065 A>G	
11657 C>A		15214 G>A	*10313 G>A*	6302 A>G	
11813 A>G		†15620 T>C	*10618 A>T*	*8086 G>A*	
12505 T>C		15627 A>G	11945 C>T	8368 C>T	
13299 T>A		15639 T>A	12330 A>G	8807 G>A	
15214 G>A		15814 C>T	13299 T>A	*9401 A>G*	
15505 T>C			13620 C>T	*10014 G>A*	
15627 A>G			13971 A>T	13299 T>A	
15639 T>A			15639 T>A	14079 T>C	
15814 C>T			15814 C>T	15639 T>A	
16025 T>C			16025 T>C	15814 C>T	
				16025 T>C	

due to mapping errors at the start of the contig. Repetitive regions identified were masked in all samples.

2.6 Results (Data Set 1)

The results of this chapter are divided into two parts. In the first part, I present results obtained from phylogeographic, mutational signature, selection and mtDNA recombination analyses of 449 CTVT tumours (Data Set 1).

2.6.1 Mitochondrial copy number

MtDNA was sequenced at ~70× coverage, indicating that each CTVT cell carries approximately 470 mtDNA copies (Appendix 1: Supplementary file 2.2). This is coincident with the amount of mtDNA observed in human cancer cells (~500 copies) (Ju et al., 2015).

2.6.2 Mutational processes in CTVT mitochondria

We identified 1005 single point substitution variants (Appendix 1: Supplementary file 2.4a) and 27 short indels in the CTVT mtDNA population (Appendix 1: Supplementary file 2.5a). CTVT mtDNA somatic substitution mutations had the characteristic profile that is observed in human cancers (Ju et al., 2014), dominated by C>T and T>C mutations showing a striking strand bias (Fig. 2.5). This mutational process is probably replication-coupled, and mutations associated with this process appear to accumulate at a roughly constant rate in human cancers (Ju et al., 2014).

2.6.3 CTVT phylogeny

A maximum likelihood phylogenetic tree constructed with mtDNA sequences from CTVT, matched hosts and hundreds of additional dogs revealed that CTVT mtDNAs cluster in five distinct groups within dog mtDNA haplogroup A1 (Fig. 2.6). These data suggest that CTVT mtDNAs have at least five independent origins, demarcating five distinct horizontal transfer groups or clades. The geographic distribution and phylogenies of the five CTVT clades reveal the dynamic recent history of the CTVT lineage (Fig. 2.7a and c). Clades 1 and 2, which occur most frequently in the CTVT population analysed, both have a global distribution. Tumours that diverged early

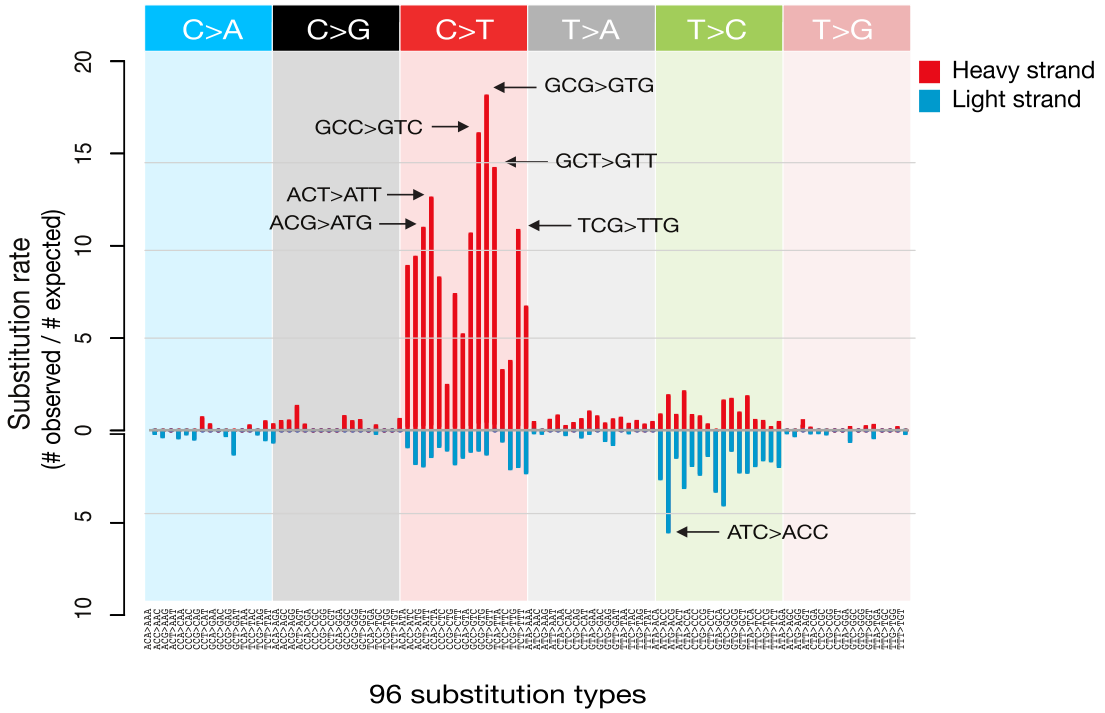


Fig. 2.5. CTVT mtDNA somatic mutation spectrum.
CTVT somatic mutations displayed by mutation type (in pyrimidine context) with 5′ and 3′ context and strand. Only conservative somatic substitutions were used to generate the spectrum (Section 2.4.9.1). Each of 96 mutation classes is displayed on the horizontal axis, with mutations occurring on the heavy strand displayed in red on the positive axis, and light strand mutations displayed in blue on the negative axis. The normalised substitution rate represents the number of observed mutations divided by the number of expected mutations, given mtDNA genome triplet content. Distinctive peaks are individually labelled.

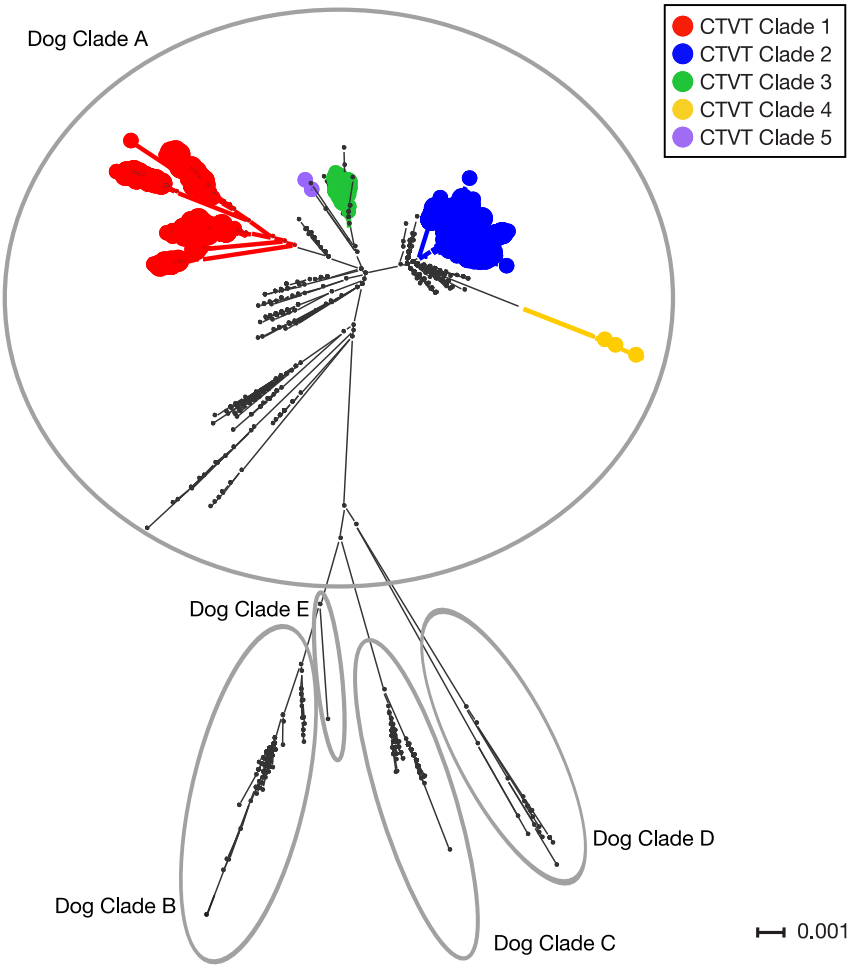


Fig. 2.6. Maximum likelihood phylogenetic tree constructed with complete mtDNA sequences from 449 CTVTs and 590 dogs (including matched hosts). Coloured and black dots represent CTVT and dog mtDNA respectively. Scale bar indicates base substitutions per site. See Appendix 1: Supplementary file 2.12 for source data.

in the clade 1 lineage occur in Russia, Ukraine, China and India, suggesting an Old World origin for this clade. CTVT samples from the Gambia appear to represent an early, divergent branch of CTVT clade 2 tumours (Fig. 2.7c).

The more recent clade 3 lineage is found in Central and South America and India, and the less frequent clades 4 and 5 occurred only in India and Nigeria respectively (Fig. 2.7). The extensive and recent global expansion detected in the CTVT lineage is consistent with signals of widespread admixture observed in worldwide populations of domestic dogs (Shannon et al., 2015), highlighting the extent to which canine companions accompanied human travellers on their global explorations.

Table 2.4 Estimated timing of clade divergence excluding potential somatic mutations (CTVT clades 1-5).

Clade	# somatic mutations (mean)	Divergence times (ybp)			
		Nuclear DNA	# cell divisions (first stage)	# cell divisions (second stage)	# mutations per year
1	18.7	911.3	204.7	1023.6	747.8
2	9.4	460	103.3	516.7	377.5
3	5	243.7	54.8.	273.8	200
4	18.7	909.9	204.4	1022.1	746.7
5	0.0	0.0	0.0	0.0	0.0

Clade 4 may have been widespread in central Asia in the past and now replaced by the more recent clades 1 and 2. CTVT tumours sampled from India represent the most divergent phylogenetic group. Two clade 4 tumours that diverged approximately 900 years ago were both sampled in Jaipur suggesting that clade 4 may have been present in this area since that time.

2.6.4 Timing of horizontal transfer events

The estimated number of somatic mtDNA mutations acquired by each clade since mtDNA capture was used to investigate the relative time since each mtDNA horizontal transfer event in CTVT. Somatic substitution analysis showed that clade 1 mtDNA carry more than double the number of somatic mtDNA mutations (22.5 mutations average) compared with clade 2 mtDNA (9.4 mutations average). Whole genome sequences of two CTVT tumours derived from clades 1 and 2 indicated that these two clades shared a common ancestor approximately 460 years ago (Murchison et al., 2014). Assuming that the clade 2 mtDNA horizontal transfer event occurred no more than 460 years ago, this analysis suggests a maximum time since mtDNA uptake of 1,097 years for clade 1, 244 years for clade 3, 1,690 years for clade 4 and 585 years for clade 5, assuming a constant somatic accumulation of mutations in CTVT mtDNA. Importantly, two additional mutation rate estimates, derived using human data (Ju et al., 2014), provided similar estimates for CTVT clade temporal origins (Section 2.4.11).

2.6.5 Tracing CTVT phylogeography

Clade 1 tumours in Central and South America share a single common ancestor that probably existed no more than 511 years ago, suggesting introduction of CTVT to

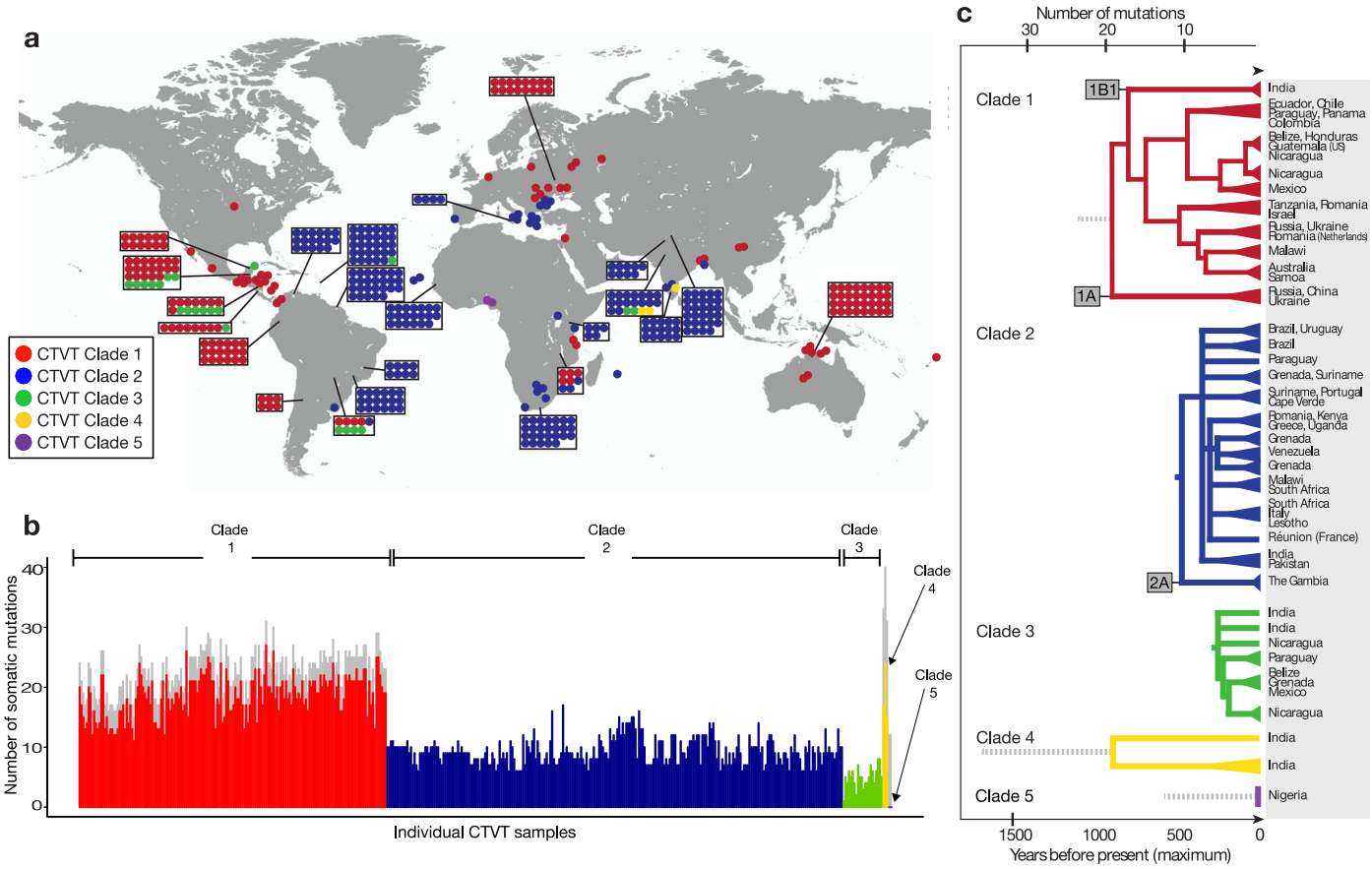


Fig. 2.7. a. Geographical distribution of clades. Each dot represents the location of CTVT sample collection, coloured by CTVT mtDNA clade. **b.** Number of somatic substitution mutations per CTVT tumour. Coloured bars indicate somatic substitutions acquired by each tumour since mtDNA capture. Grey bars indicate potential somatic substitutions; thus the early somatic or rare germline status of these variants is unknown. **c.** Simplified representation of maximum likelihood phylogenetic trees for each clade. Trees illustrate nodes with bootstrap support >60, and shaded triangles represent coalescence of individual branches within each group. Basal trunks leading to haplogroups 1A, 1B1, and 2A are labelled. Two tumours were collected in the United States and the Netherlands from dogs imported from Guatemala and Romania, respectively. Discontinuous grey lines represent contributions of potential somatic substitutions. Assuming a constant accumulation of mutations within and between clades, approximate number of somatic mutations and estimated timing is shown. Figure adapted from Strakova et al. (2016). See Appendix 1: Supplementary file 2.13 for source data.

Table 2.5 Estimated timing of clade divergence including potential somatic mutations (CTVT clades 1-5)

Clade	# somatic mutations (mean)	Divergence times (ybp)			
		Nuclear DNA	# cell divisions (first stage)	# cell divisions (second stage)	# mutations per year
1	22.5	1097.1	246.5	1232.4	900.2
2	9.4	460	103.3	516.7	377.5
3	5	243.7	54.8	273.8	200
4	34.7	1689.9	379.6	1898.2	1386.7
5	12	585.0	131.4	657.1	480

the Americas coincident with colonial contact. Similarly, this analysis suggests a single introduction of CTVT to Australia after European arrival (maximum 116 years ago; Section 2.4.11). The distribution pattern and timing of clade 2 suggest a rapid burst of expansion from a single place and that this clade may have been transported between continents along maritime trans-Atlantic and Indian Ocean trade routes.

2.6.6 MtDNA haplotype of the CTVT founder dog

This analysis indicates that the original CTVT mtDNA haplotype, present in the dog that first spawned CTVT, has been replaced and is not detectable in any of the analysed tumours. Moreover, as mtDNA in CTVT has been acquired by horizontal transfer, analysis of CTVT mtDNA could not cast light on the timing of CTVT's origin.

2.6.7 Selection

We searched for evidence of mtDNA functionality in CTVT cells by examining CTVT mtDNA for signals of negative selection. If present, negative selection would be expected to operate on mtDNA to prevent homoplasmy of deleterious mutations. Consistent with this prediction, the VAF of nonsense substitutions (p=0.00019; Fig. 2.8a) and frameshift indels (p=3.03x10^{-05}; Fig. 2.8b) was significantly lower than VAF for other substitutions and indels. Furthermore, dN/dS for somatic mtDNA mutations in CTVT showed significant deviation from neutrality (namely dN/dS $<$1) both for nonsense (0.187, p=1.02x10^{-07}) and missense (0.748, p=4.18x10^{-03},) mutations (Fig. 2.9).

These findings provide evidence for the activity of negative selection operating to

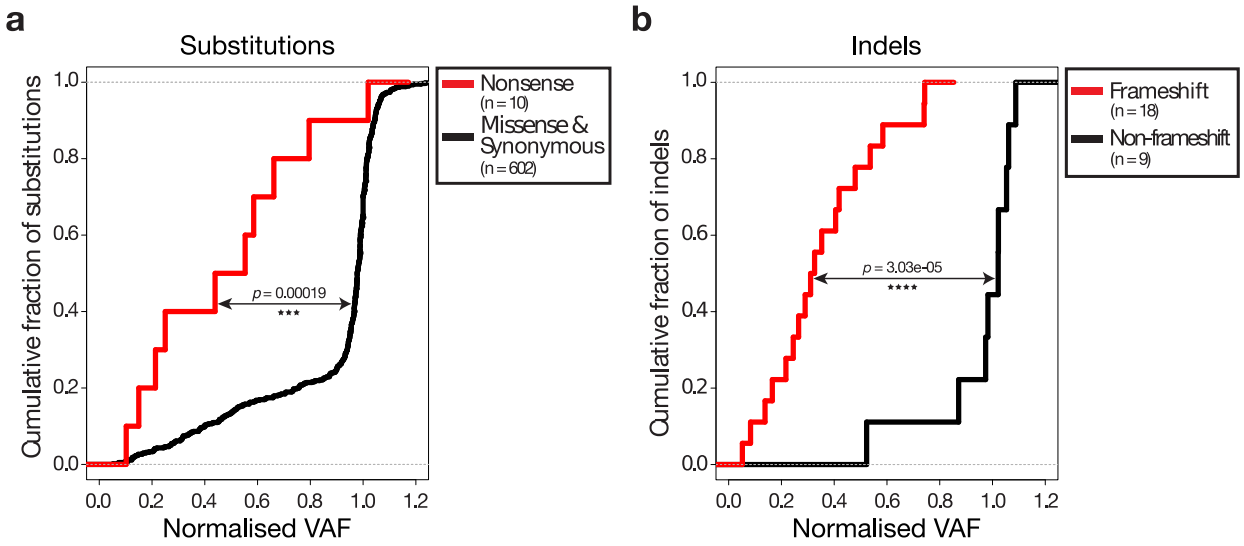


Fig. 2.8. Cumulative variant allele fraction distribution functions for gene-disrupting **a.** substitutions and **b.** indels. *P*-values were calculated using two-sample Kolmogorov-Smirnov tests. Indel VAF values were normalised to take account of host contamination. Figure adapted from Strakova et al. (2016).

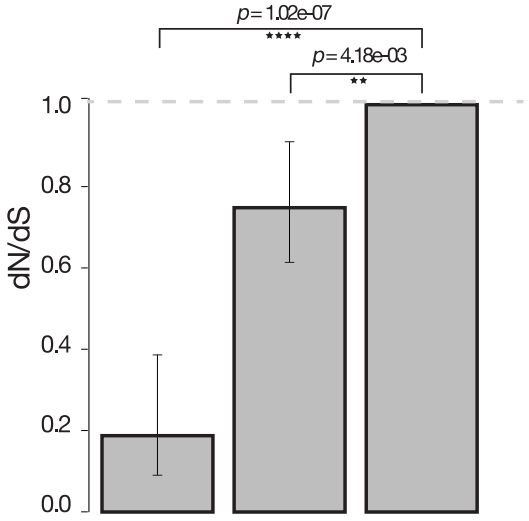


Fig. 2.9. dN/dS for somatic nonsense and missense substitutions. P-values were calculated using a likelihood ratio test with parameters estimated using a Poisson model. Error bars indicate 95 percent confidence intervals. ** P<0.01, **** P<0.0001

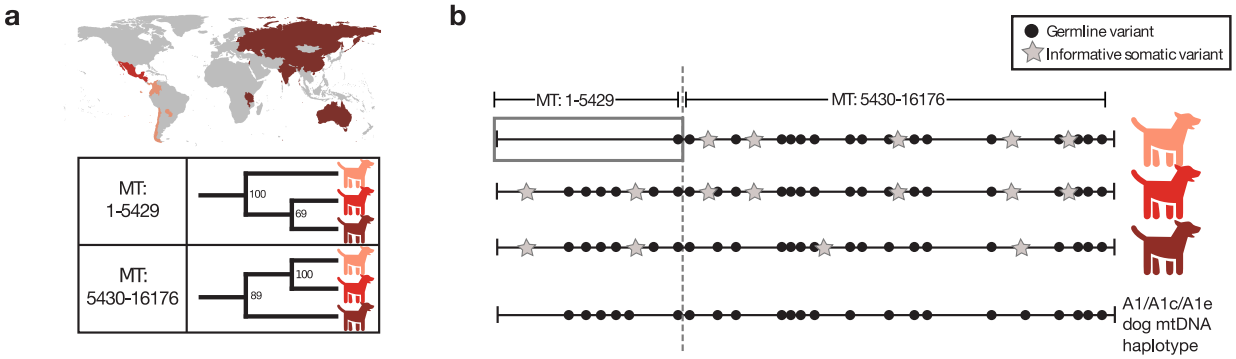


Fig. 2.10. Ancient mtDNA recombination in CTVT
a. Maximum likelihood phylogenetic trees constructed using clade 1 CTVT mtDNA segments MT:1–5429 and MT:5430–16176. Three clade 1 mtDNA haplogroups are represented by coloured dog silhouettes, and their geographical distributions are colour-coded on the map. Bootstrap values were calculated from 100 iterations. See Appendix 1: Supplementary file 2.14 for source data. **b.** Simplified haplotype diagrams for clade 1 CTVT mtDNAs derived from groups shown in **a**. A1/A1c/A1e dog haplotype represents the donor mtDNA that founded clade 1. The regions putatively replaced by recombination is outlined with a grey box. Figure adapted from Strakova et al. (2016).

prevent homoplasmy of deleterious coding variants and preserve mtDNA function in CTVT. Together with evidence of reduced VAF for truncating mtDNA mutations in human cancers (Ju et al., 2014; Stewart et al., 2015; Yuan et al., 2017), this suggests that, in some cancers, functional mtDNA is required to drive cancer.

2.6.8 Ancient mtDNA recombination in CTVT

Given the opportunity for multiple distinct mtDNA haplotypes to coexist in CTVT cells, we searched for evidence of mtDNA recombination in CTVT using the recombination-detection algorithms 3seq (Boni et al., 2007) and SiScan (Gibbs et al., 2000). These algorithms detected evidence for mtDNA recombination in CTVT clade 1, detecting recombination breakpoints near positions MT:5430 and MT:16176. Maximum likelihood phylogenetic trees constructed using segments MT:1–5429 and MT:5430–16176 derived from CTVT clade 1 mtDNAs showed distinct topologies (Fig. 2.10a). Further inspection of clade 1 mtDNA haplotypes suggested that a recombination event replaced MT:1–5429 in a clade 1 mtDNA haplotype (Fig. 2.10b) that diverged from Central American clade 1 CTVTs and that subsequently colonised regions in South

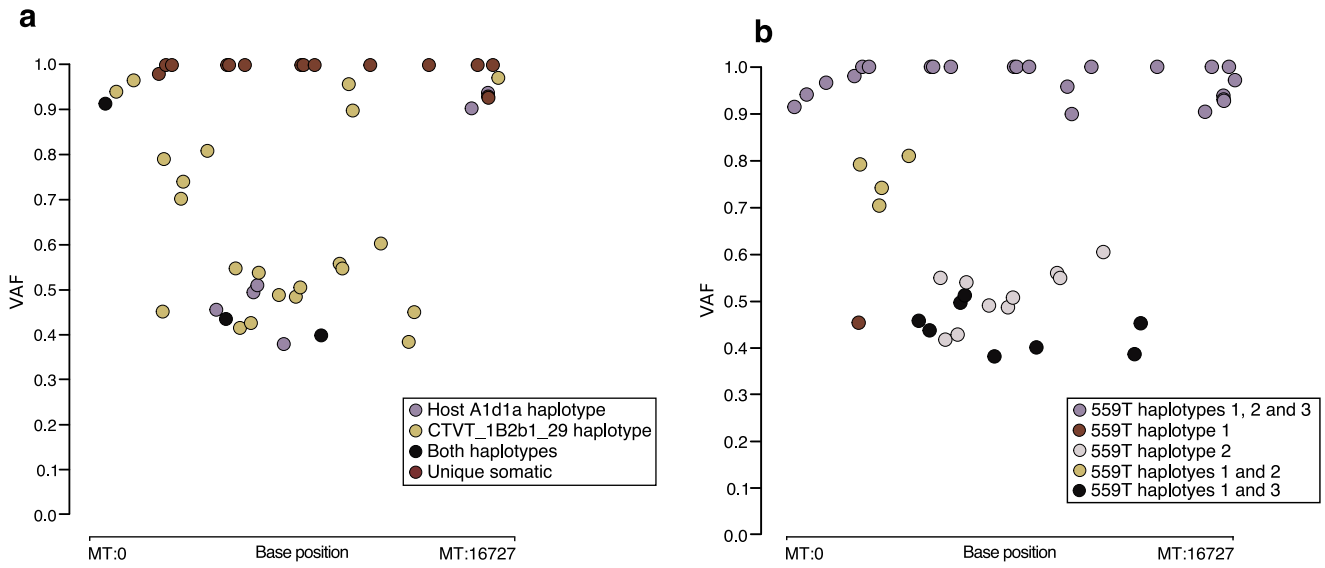


Fig. 2.11. Variant allele fraction plots illustrating mtDNA recombination in tumour 559T and showing **a.** the dog mtDNA haplotype (Host A1d1a_1) and CTVT clade 1 haplotype (CTVT_1B2b1_29) involved in recombination and **b.** the three most common recombinant haplotypes present within the 559T mtDNA population, inferred from the phasing of long-read data.

and Central America (Panama, Ecuador, Chile, Paraguay and Colombia). These data provide evidence of an mtDNA recombination event in an ancestral CTVT lineage.

2.6.9 Recent mtDNA recombination in CTVT

Evidence of more recent mtDNA recombination was assessed by examining outliers on CTVT mtDNA phylogenetic trees. This analysis identified 559T, a CTVT tumour derived from a male golden retriever in Nicaragua. Further investigation of mtDNA in 559T revealed what appeared to be a CTVT clade 1 mtDNA haplotype (CTVT_1B2b1_29) superimposed upon a dog mtDNA haplotype (A1d1a_1), neither of which resembled the mtDNA haplotype found in normal tissues from the matched host dog, 559H (B1_1 haplotype; Fig. 2.11).

Phasing of mtDNA variants in 559T using long sequence reads indicated the presence of at least three distinct mtDNA haplotypes in this tumour, each representing a recombination product that appeared to be derived from mtDNA haplotypes CTVT_1B2b1_29 and A1d1a_1. These data suggest that a tumour antecedent of 559T

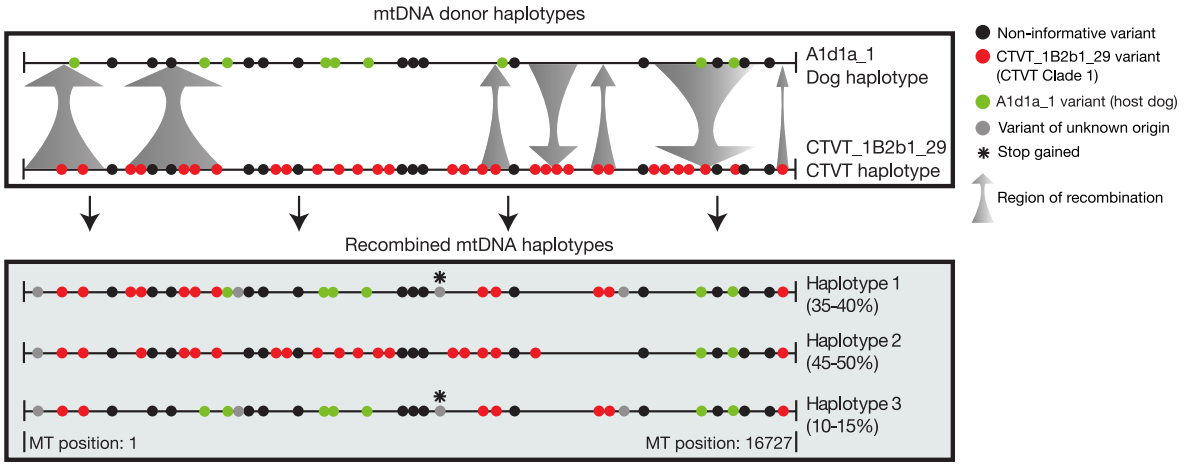


Fig. 2.12. Recent mtDNA recombination in CTVT
Recombination detected in CTVT sample 559T from Nicaragua. The estimated percent contribution of each recombined haplotype to the mtDNA population within 559T CTVT cells is shown, and grey shaded arrows indicate putative regions of mtDNA recombination. Figure adapted from Strakova et al. (2016).

captured haplotype A1d1a_1 mtDNA from its host. Recombination was initiated between mtDNA haplotypes CTVT_1B2b1_29 and A1d1a_1, and cells containing these recombination products were passed to host 559H. Alternatively it is possible that 559H received a mixture of both normal and CTVT cells from its CTVT donor animal, and mtDNA capture and recombination occurred within 559H. It must also be mentioned that the A1d1a_1 haplotype resembles the CTVT clade 3 donor haplotype. The possibility that the recombination observed in sample 559T involved horizontal transfer between CTVT clade 1 and clade 3 tumours that occurred within the same animal cannot be excluded.

2.7 Results (Data Set 2)

In the second part, I present results from phylogeographic and recombination analyses applied to an extended data set of 640 CTVT samples.

2.7.1 mtDNA recombination

A further instance of recent mtDNA recombination was identified in 1315T, a CTVT tumour derived from a female dog in Granada, Nicaragua. The donor haplotypes observed in 1315T were identical to the donor haplotypes described earlier in 559T (Sec-

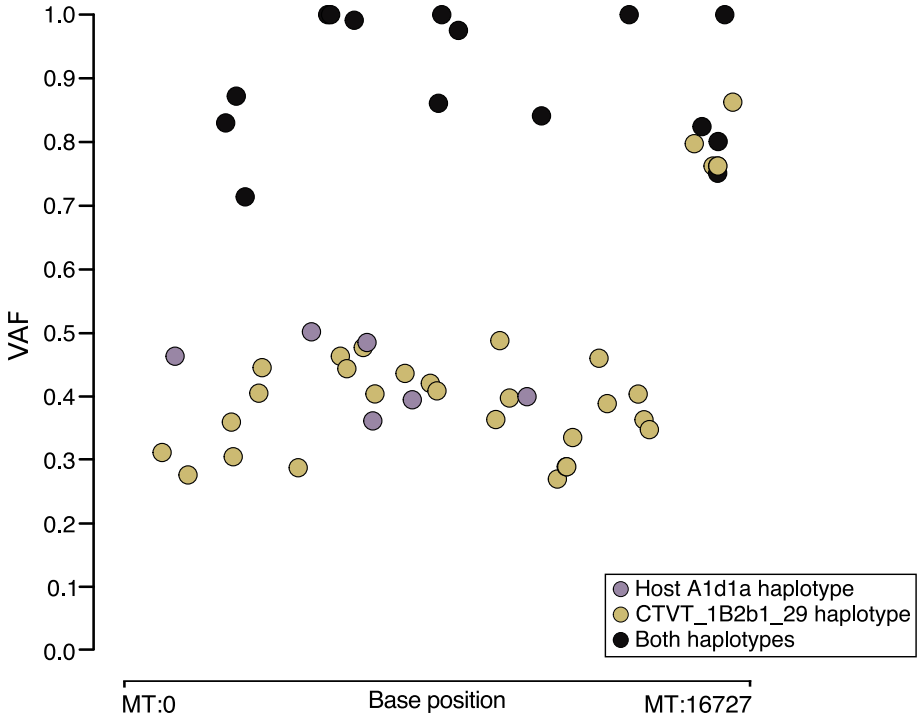


Fig. 2.13. Variant allele fraction plot for recombinant CTVT sample, 1315T. MtDNA recombination in 1315T involved a dog mtDNA haplotype (Host A1d1a_1) distinct from the 1315H matched host haplotype (A1a1a_28) and a CTVT clade 1 haplotype (CTVT_1B2b1_29). These are the same donor mtDNA haplotypes observed in 559T.

tion 2.6.8): a likely CTVT clade 1 mtDNA haplotype (CTVT_1B2b1_29) interspersed with a dog mtDNA haplotype (A1d1a_1), distinct from the mtDNA haplotype found in normal tissues from the matched host dog, 1315H (A1a1a_28 haplotype).

2.7.2 Recent mtDNA horizontal transfer

Two samples out of 640 were identified in which a second horizontally transferred host mtDNA haplotype could be detected. In sample 1281T, a third mtDNA haplotype at ~34% from the B1 haplogroup (along with the tumour haplotype and contaminating host mtDNA haplotype) was detected (Fig. 2.14). This is the first horizontally transferred mtDNA not derived from the dog Clade A haplogroup. In sample 349T2, a host A1a1a haplotype was detected at ~15% (Fig. 2.14).

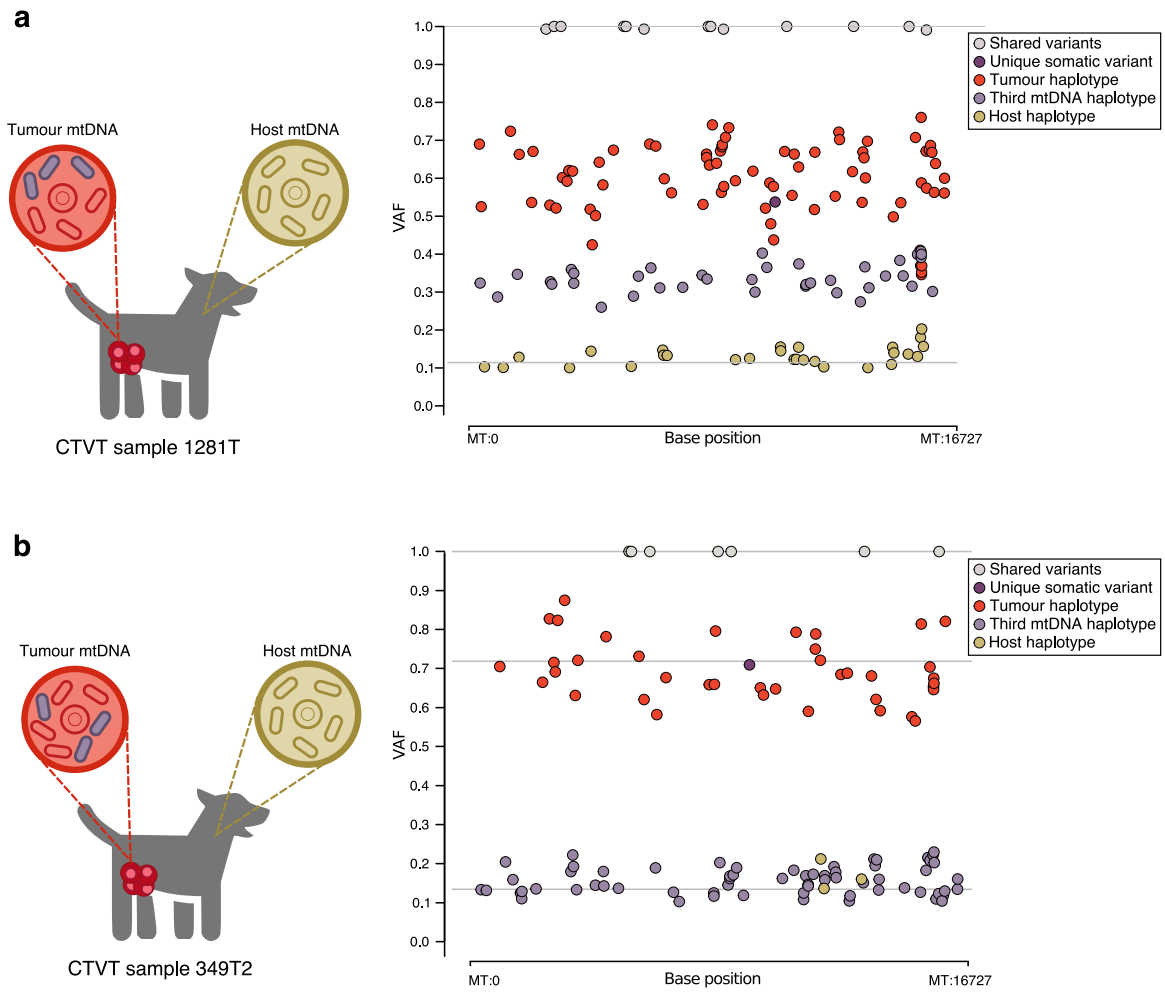


Fig. 2.14. Variant allele fraction plots for CTVT samples **a.** 1281T and **b.** 349T2 in which three different mtDNA haplotypes were identified.

Clade	A1d1 horizontal transfer group	Location	Sample(s)	Haplogroup
5	1	Ibadan, Nigeria	627T1, 628T1	A1d1
8	2	Corozal District, Belize	464T, 738T, 739T, 740T, 782T, 839T1, 1340T	A1d1a1
9	3	San Juan del Sur, Nicaragua	94T	A1d1a
10	3	Granada, Nicaragua	559T, 1315T	A1d1a
11	4	Granada, Nicaragua	540T, 550T, 551T, 555T, 564T, 572T, 1283T, 1284T, 1288T, 1290T, 1297T, 1298T, 1304T, 1311T	A1d1a1
12	5	Kombo, The Gambia	1172T	A1d1a1
13	6	Grenada	342T	A1d1a1
14	7	Asuncion, Paraguay	363T, 364T, 365T, 366T	A1d1a
15	8	Jaipur, India	313T, 411T, 980T	A1d1a
16	9	Los Andes, Chile	958T	A1d1a1
17	10	Manizales, Colombia	1539T	A1d1a
18	11	Manizales, Colombia	1353T	A1d1a1
19	12	Mexico	369T*	A1d1a1

Table 2.6 Summary of recurrent horizontal transfers of A1d mtDNA haplogroup Individual horizontal transfer events were inferred based on comparison with the nuclear CTVT phylogenetic tree. Sample 369T is marked with an asterisk as this sample was not included in the exome sequencing sample set.

2.7.3 Recent and repeated horizontal transfer of a single mtDNA haplotype

Comparison of CTVT phylogenetic trees generated using mitochondrial and nuclear genomic data (Fig. 2.15) unexpectedly showed that CTVT clade 3 tumours were not monophyletic, as was suggested during analysis of Data Set 1. Instead, nuclear data implied that what appeared to be a single clade was in fact 13 independent horizontal transfers involving the same mitochondrial donor haplogroup (A1d1). This is the only instance observed so far in CTVT mitochondria where a specific haplogroup has undergone repeated horizontal transfer.

2.7.4 Expanded CTVT phylogeny

Analysis of 640 complete CTVT mitogenomes confirmed the previous CTVT phylogenetic structure and revealed two new clades, 6 and 7, identified in Armenia and Mexico, respectively (Fig. 2.15). To date, twenty horizontal transfer events involving distinct mtDNA haplotypes have been detected in CTVT. In seventeen of these cases,

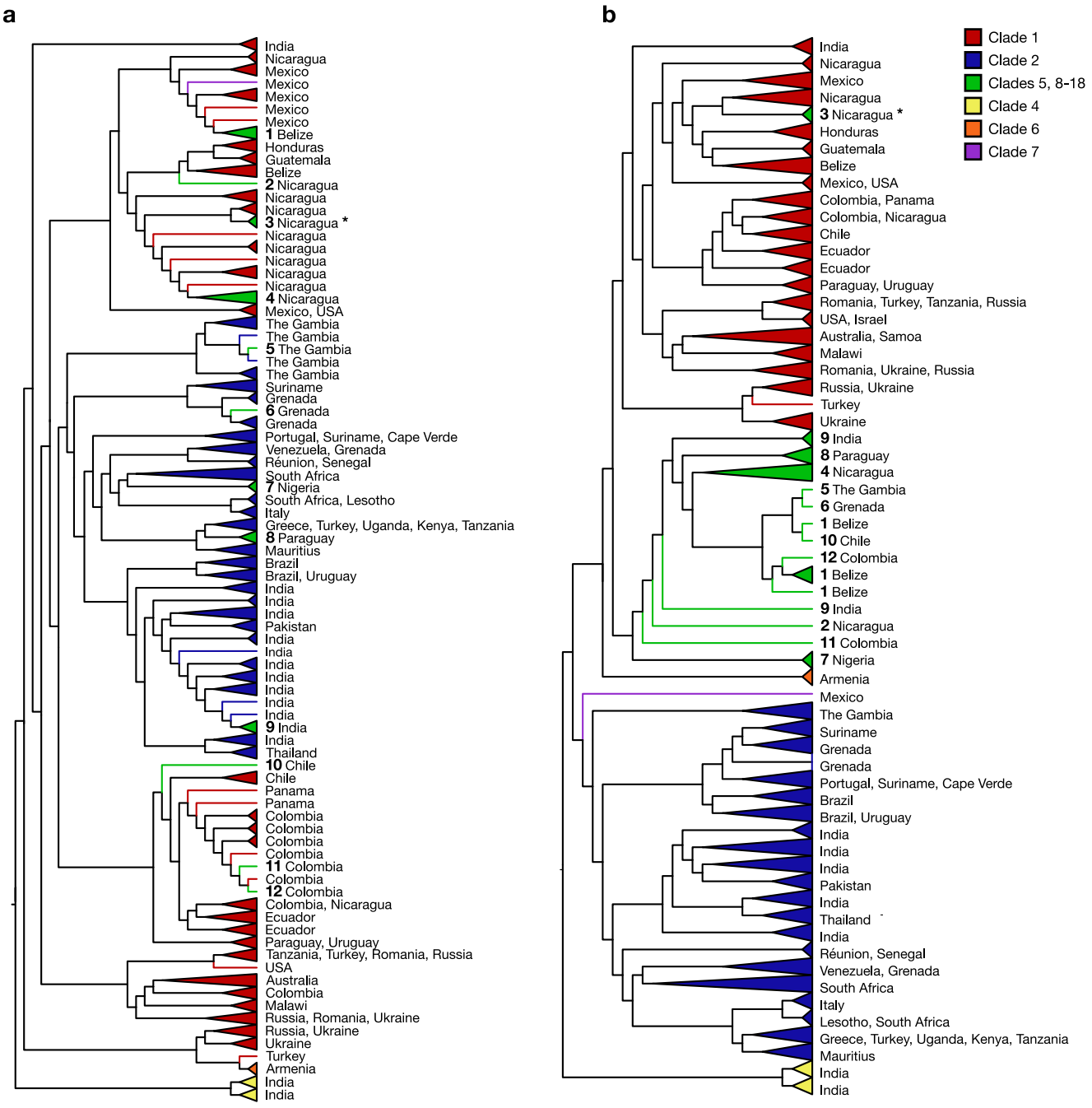


Fig. 2.15. Schematic of maximum likelihood phylogenetic trees generated from **a.** nuclear and **b.** mitochondrial CTVT sequence data. Phylogeny inferred from whole-exome data was constructed by Adrian Baez-Ortega. Triangles, coloured according to CTVT clade, represent the collapsed tree branches of phylogenetic tumour groups. Each tip is a single tumour sample. Sample locations are labelled. Numbers beside tumour groups count independent horizontal transfers of the A1d1 mtDNA haplogroup. An asterisk (*) is used to label the phylogenetic tumour group made up of the samples with recombinant mtDNA, 559T and 1315T.

Table 2.7 Estimated timing of clade divergence excluding potential somatic mutations (CTVT clades 5-19)

Clade	# somatic mutations (mean)	Divergence times (ybp)			
		Nuclear DNA	# cell divisions (first stage)	# cell divisions (second stage)	# mutations per year
5	0.0	0.0	0.0	0.0	0.0
6	1	48.78	54.64	10.95	40
7	0.0	0.0	0.0	0.0	0.0
8	2.86	139.37	156.13	31.29	114.29
9	0	0.0	0.0	0.0	0.0
11	0.5	24.39	27.32	5.48	20
12	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0
14	1.25	60.98	68.31	13.69	50
15	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0
17	0.0	0.0	0.0	0.0	0.0
18	0.0	0.0	0.0	0.0	0.0
19	0.0	0.0	0.0	0.0	0.0

the captured mtDNA has become fixed: CTVT clades 1, 2, 4, 6, and 7 along with 12 independent captures of a specific A1d1 haplotype, while in three cases, 559T and 1315T (Clade 10; Table 2.6), 1281T and 349T2, the mtDNA appears to have been captured recently and has not become fixed. It is possible that mtDNA horizontal transfer takes place in CTVT cells more frequently than estimated in this data set but that the majority of these events do not result in the fixation of the transferred mtDNAs and so, are not observed.

2.7.5 NuMTs

Analysis of germline NuMTs confirmed that NuMTs that form part of the CanFam3.1 assembly did not have an impact on the variant analysis performed in this chapter (Section 2.4.15). The average MT genome coverage from this analysis was 0 (Fig. 2.16), indicating that the NuMTs known to be present within CanFam3.1 are at insufficient copy number and/or are too divergent to map to the MT reference genome using the alignment parameters used in this chapter.

Table 2.8 Estimated timing of clade divergence including potential somatic mutations (CTVT Clades 5-19)

Clade	# somatic mutations (mean)	Divergence times (ybp)			
		Nuclear DNA	# cell divisions (first stage)	# cell divisions (second stage)	# mutations per year
5	10	0	0	0	
6	11	48.78	54.64	10.95	40
7	0	0	0	0	0
8	2.86	139.37	156.13	31.29	114.29
9	4	195.12	218.58	43.81	160
11	0.5	24.39	27.32	5.48	20
12	1	48.78	54.64	10.95	40
13	1.0	48.78	54.64	10.95	40
14	2.25	109.76	122.95	24.64	90
15	0.67	32.52	36.43	7.30	26.67
16	5	243.90	273.22	54.76	200
17	0.0	0.0	0.0	0.0	0.0
18	0.0	0.0	0.0	0.0	0.0
19	1	48.78	54.64	10.95	40

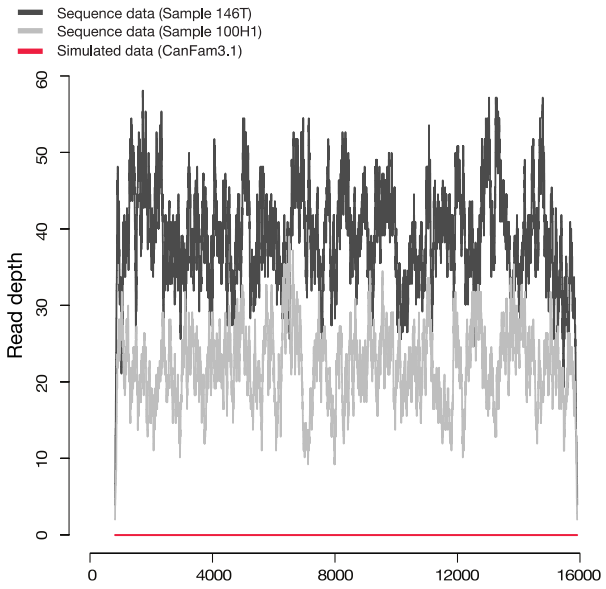


Fig. 2.16. NuMT DNA sequence contribution
Sequence read depth across the MT genome for representative CTVT (146T) and host (100H1) samples sequenced in this chapter to ~0.3× whole genome average coverage. Sequence read depth for simulated reads from CanFam3.1 (excluding the MT chromosome) is also shown in red; reads were simulated to ~0.3× whole genome average coverage.

2.8 Discussion

2.8.1 Horizontal mtDNA transfer in CTVT

Following previous work by Rebbeck et al. (2011), this analysis of CTVT mtDNA shows that independent mtDNA horizontal transfer has occurred at least twenty times, with eighteen of these transfers defining tumour clades whose global distributions track two millennia of CTVT spread and trace the movement of dogs around the world.

Given the findings of this chapter, an interesting direction for future research would be understanding the mechanisms underlying mitochondrial behaviour in CTVT. In particular, an interesting direction for future work would be understanding the mechanism of intercellular horizontal mtDNA transfer in CTVT and the molecular signals driving this process. For example, it is not yet known whether naked mitochondria are transferred from host donor cells to recipient CTVT cells or whether mtDNA is packaged within extracellular vesicles and transferred to CTVT cells along with other host 'cargo' (e.g. mRNA, proteins, lipids; Sansone et al. 2017).

2.8.2 Selection pressures in CTVT mitochondria

Negative selection has operated in CTVT to maintain mtDNA integrity by preventing the accumulation of gene-disrupting (nonsense and missense) mutations in captured mtDNA. This finding demonstrates that maintenance of functional mtDNA is important for the biology of CTVT. This supports the importance of functional mtDNA in cancer and reveals novel biological mechanisms that have operated in an ancient mammalian somatic cell lineage.

Despite showing that negative selection operates in CTVT mitochondrial genomes, our findings could also suggest that this activity is not sufficient to maintain functional mtDNA over the course of thousands of years and, eventually, recruitment of host mtDNA may provide a selective advantage to cells. This could explain why the earliest mtDNA capture identified in this data set took place ~2,000 years ago, 9,000 years after the estimated emergence of CTVT (Murchison et al., 2014). However, the possibility that mtDNA horizontal transfer may have taken place much earlier in unsampled CTVT lineages cannot be excluded.

2.8.3 Frequent horizontal transfer of a specific mtDNA haplotype

One of the most exciting results of this analysis is the finding of repeated and specific mitochondrial uptake of a host mtDNA haplotype. It had previously been observed that clade 3 tumours did not form a monophyletic group in the context of normal dog haplotypes (Appendix 1: Supplementary file 2.12). However, this could be explained if somatic mutations arising in clade 3 samples occurred at sites also found in germline haplotypes closely related to the clade 3 donor haplotype. This seemed plausible given that we found germline and somatic mtDNA mutational processes are equivalent (Section 2.6.2) and clade 3 was initially interpreted as a single clonal lineage.

Instead, nuclear data has indicated repeated horizontal transfer of mtDNA from a specific dog mtDNA haplogroup. A number of different hypotheses could explain this observation. It could indicate that A1d1 mtDNA haplotypes are recruited by CTVT cells via horizontal transfer more frequently than other haplotypes. It is also possible that horizontal transfer of mtDNA occurs frequently in CTVT and all dog mitotypes are equally likely to undergo horizontal transfer. However, the majority of transferred mtDNAs and do not persist. A1d1 mtDNA haplotypes might possess a proliferative or survival advantage and become fixed in the mitochondrial population.

To study the timing of the newly detected horizonal transfer events we again estimated the number of somatic mtDNA mutations accumulated by each clade since host mtDNA uptake. This analysis revealed that mtDNA derived from horizontal transfers involving A1d1 haplotypes (clades 5, 8-19) carry few somatic mutations (Tables 2.7 and 2.8) and all A1d1 horizontal transfers occurred relatively recently, most within the last decades (Tables 2.7 and 2.8).

Given that CTVT has been routinely treated with vincristine sulphate, a cytotoxic microtubule inhibitor, since the 1980s (Amber et al., 1990; Section 1.2.1.4) and the evidence that chemotherapeutic agents can enhance mitochondrial horizontal transfer *in vitro* (Moschoi et al.,2016; Section 2.3.5), combined with the recent and repeated recruitment of A1d1 mtDNA haplotypes from host cells, it is worth considering what role chemotherapy related effects might play in this phenomenon. For one, it could be that A1d1 haplotypes specifically are more likely than other haplotypes to undergo horizontal transfer when CTVT is exposed to chemotherapy. Alternatively,

chemotherapy treatment could increase the opportunity of horizonal transfer for all mtDNA haplotypes but A1d1 haplotypes may have a selective advantage over incumbent CTVT mtDNA haplotypes. Another hypothesis would be that vincristine does not make a difference to the rate of horizontal transfer in CTVT cells but that after uptake, A1d1 haplotypes are more like to persist in vincristine exposed CTVT cells. However, testing these hypotheses is beyond the scope of the present dissertation.

Lastly, the observation of multiple mtDNA horizontal transfer events in Data Set 1 at first suggested that mtDNA capture from hosts might be a positively selected adaptive mechanism (Spees et al., 2006; Tan et al., 2015) and that capture of host mitochondria could offer a fitness advantage to the cell (Rebbeck et al., 2009). However, we cannot exclude the possibility that the acquired mtDNA haplotypes became homoplasmic as a result of selfish advantage, e.g. increased replicative efficiency relative to other haplotypes, which might not be of benefit to the CTVT cell or that CTVT cells have simply acquired host mtDNA via drift.

2.8.4 mtDNA Recombination

Ancient mtDNA recombination occurred in a clade 1 mitotype that subsequently colonised the Pacific coast of South America and there is evidence for the occurrence of complex, recent recombination in at least two tumours, possibly to repair damaged mtDNA. Further studies might explore the extent of mtDNA recombination on a wider scale in CTVT and in cancer in general, including human cancers.

Two tumour samples were identified in which A1d1a and CTVT_1B2b1_29 haplotypes have undergone recombination. The most likely explanation for these data is that a tumour antecedent of 559T and 1315T captured an A1d1a_1 mtDNA haplotype from its host. Recombination was initiated between mtDNA haplotypes CTVT_1B2b1_29 and A1d1a_1, and cells containing these recombination products were passed to hosts 559H and 1315H. Less plausible scenarios are that

(i) Both 1315H and 559H received a mixture of normal and CTVT cells from the same CTVT donor animal; mtDNA capture and recombination was initiated independently within both 559H and 1315H.

(ii) Both 559H and 1315H were co-infected with CTVT clade 1 and clade 3 tumours; the observed recombination in samples 559T and 1315T resulted from horizontal transfer between these clade 1 and clade 3 mtDNAs.

(iii) Dog 1315H was directly infected with tumour cells from dog 559H; a fraction of cells from 559T carrying mtDNA recombination products not detected or sampled in this study were transmitted and seeded 1315T.

(iv) Dog 559H was directly infected with tumour cells from dog 1315H; a fraction of cells from 1315T carrying mtDNA recombination products not detected or sampled in this study were transmitted and seeded 559T.

The latter two scenarios are unlikely given that clinical records show dog 559H was neutered at the time of sampling (July 2013) and so, is unlikely to have mated successfully with 1315H subsequent to that.

At the time of sampling (March 2015), dog 1315H's tumour was roughly 4 inches in size. Unless there is a large variation in tumour growth rates, tumour 1315T is unlikely to have resulted from an infection >2 years prior. Given this time frame, 1315T is unlikely have resulted from direct infection by cells from 559T and unlikely to have caused tumour 559T by direct transmission of cells from 1315T.

Although there was no evidence of mtDNA recombination in CTVT beyond that described, it cannot be excluded that recombination is more widespread in CTVT mtDNA than detected. It is possible, therefore, that the phylogenetic, mutation rate and selection analyses presented in this chapter have been influenced by an undetected recombination signal. However, the following observations suggest that if such a signal is present in this data set, it is at a low level:

(i) the presence of a set of clade-specific markers in all non-recombining CTVT mtDNAs,

(ii) the absence of distinctive phylogenetic outliers (beyond 559T and 1315T),

(iii) the very low frequency of back-mutation,

(iv) the strong somatic signal identified in the CTVT mtDNA mutational spectrum, and

(v) the failure of recombination-detection algorithms to detect further recombination.

Detection of mtDNA recombination in ordinary cancers using the approaches described in this chapter poses significant challenges; clonal inheritance of mitochondria means that most recombination events would be 'silenced' by a lack of biomarkers allowing different MT molecules to be distinguished.

The mechanism by which different mtDNA molecules are able to interact within the cell and the nature of the signals that trigger onset of mtDNA recombination remains unclear. It is interesting to observe that a truncating nonsense mutation in *COX3* (9064 G>A) was found in some of the recombinant 559T haplotypes (Fig. 2.12). Further analysis may determine if DNA damage signalling is involved, as previously suggested (Thyagarajan et al., 1996). Many intriguing open questions remain including whether mtDNA recombination in CTVT occurs sporadically or is triggered, whether mtDNA recombination be triggered by mitochondrial diversity within the cell.

2.8.5 NuMTs

While it has been shown that NuMTs present in the CanFam3.1 assembly are unlikely to have impacted the mutation calling performed in this chapter (Section 2.7.5), it remains possible that NuMTs not captured in the assembly could confound the variant analysis. The following observations argue against the possibility that NuMT-derived variants have had a significant impact on tumour variant calling:

(i) CTVT is a clonal lineage, therefore somatically acquired NuMT-derived variants would be expected to appear as low level heteroplasmic variants across all tumours within a phylogenetic group; variants with these features were not observed,

(ii) The mutation spectrum observed in CTVT mtDNA has the distinctive profile characteristic of the known somatic mtDNA mutational process (Ju et al., 2014). As this mutational process is specific to cytoplasmic mtDNA, this finding suggests that the majority of variants within the somatic variant set are of cytoplasmic origin.

In addition, I have not directly assessed the presence or frequency of mitochondrial-nuclear DNA fusions in CTVT. The observation of mtDNA horizontal transfer and recombination implies the movement of mtDNA within intracellular and/or extracellular compartments; this could allow for a high frequency of contact between

mitochondrial and nuclear DNA. A recent study by Ju et al. (2015) observed that nuclear transfer of mtDNA is unexpectedly frequent in human somatic cells. Altogether, it is likely that novel NuMTs, which diverge from those in the canine genome, have been introduced into the CTVT genome.

One interesting possibility for future analysis would be assessing the presence and frequency of somatic NuMTs in the CTVT genome. Some of these mito-nucear fusion events may have taken place early in CTVT's evolutionary history, even before the earliest mtDNA horizontal transfer event ~2,000 years ago, and involved the nuclear capture of ancient mitochondrial dog haplotypes.

2.8.6 Summary

The ability of CTVT cells to uptake mitochondria from host stromal cells represents an emerging paradigm of cell-to-cell signalling in cancer. All cancers manipulate their host environment, but transmissible cancers, and in particular CTVT, have had much more time to do this. Recombination and horizontal mtDNA transfer may be considered as specialised adaptations that have helped CTVT co-opt its host niche and one of the ways in which this tumour manipulates its host. On the whole, the findings in this chapter show how CTVT's metastatic spread through a global host population can provide unique insights into evolutionary processes operating in cancer.

Chapter 3

Reconstructing the spatiotemporal origin of CTVT using ancient and modern genomes

3.1 Chapter abstract

CTVT has persisted in the dog population for thousands of years and first arose in an ancient population of dogs (Murchison et al., 2014; Murgia et al., 2006; Rebbeck et al., 2009). Genome-wide and SNP-base analyses of germline variation in the CTVT genome showed that the founder's closest modern relatives are the Arctic spitz breeds, such as Alaskan malamutes and huskies (Decker et al., 2015; Murchison et al., 2014). Analysis of two CTVT whole-genomes also suggested that the disease is likely to have arisen in a genetically isolated population of early dogs (Murchison et al., 2014). The mitochondrial CTVT phylogeny, described in Chapter 2, supports a model whereby CTVT originated in Central or Northern Asia. The goal of this chapter was to consolidate these findings and to further define the ancestry and origin of the CTVT founder animal as well as the timing of CTVT emergence.

Each CTVT tumour contains a living ancient genome: the long-dead CTVT founder dog's DNA has been preserved largely intact through the millennia by the cancer that its somatic cells spawned. Moreover, a recent boom in ancient sequencing studies has made the first genome-scale ancient canid DNA data sets available (Botigué et al., 2017; Frantz et al., 2016; Skoglund et al., 2015)and this prompted us to consider a new approach to uncovering the origin of CTVT: paleogenomics. This idea lead to a collaborative project with Greger Larson's lab at the Research Laboratory for Archaeology and History of Art (University of Oxford) and led by Laurent Frantz on the ancient DNA side. Initially, this study was directed at identifying the ancient ancestry of the CTVT founder individual; however, our findings provoked further

questions about the ancestral population of dogs to which the founder individual belonged and the scope of the work changed to address these questions (Section 3.3.4).

We sequenced 71 mitochondrial and 7 nuclear genomes from ancient North American and Siberian dog remains from time frames spanning ~9,000 years. This revealed that the CTVT founder was closely related to a population of pre-contact dogs (PCD) that were once widespread across North America and present in the Americas prior to arrival of European colonists. Analysis showed that this dog population were not derived from North American wolves but were, instead, the descendants of early Asian dogs introduced to the Americas during human migrations between 17,000 and 13,000 years ago. This dog population remained in North America until its almost complete disappearance following the arrival of Europeans, with minimal genetic contribution from pre-contact dogs observed in modern dog populations. Surprisingly, this implies the closest extant vestige of native American dogs is the canine transmissible venereal tumour.

Finally, in order to estimate CTVT's temporal origin, we sequenced the genomes of two CTVT tumours obtained from a non-experimental direct transmission and used these to calibrate a tumour specific mutation rate. I estimated that the CTVT founder dog lived up to 8,225 years ago, a time frame compatible with the possibility that CTVT arose in a dog population in North America.

Overall, this chapter provides new insights into when and where one of the oldest known cancers first appeared along with answering broader questions about the past population dynamics of early dogs.

3.2 Publications associated with Chapter 3

The majority of the results and analyses of this chapter were published as a Report in *Science* on the 6th July 2018, of which I am co-first author along with Angela Perri, Evan-Irving Pease, Kelsey Witt and Anna Linderholm (Ní Leathlobhair et al. (2018); see Appendix 2 for the complete reference).

The somatic variation analysis presented in Section 3.10 formed the basis of a Poisson

Bayesian model estimation of the CTVT time-of-origin reported in Baez-Ortega et al. (2018), currently under review (see Appendix 2 for the complete reference).

3.3 Introduction

While the focus of this thesis up until now has been CTVT, in this chapter we shift focus somewhat in order to address questions related to the host population in which CTVT first established its parasitic niche. Here, I will introduce findings that contextualise results presented later in this chapter related to the initial CTVT host population. Since I will later demonstrate that dogs were not domesticated in the New World, I first raise the issue of the timing of human migrations in the Americas since people were likely responsible for the introduction of dogs. I then give a brief synopsis of the history of dogs and other canids in the Americas and summarise relevant methods that can be applied to genetic data for analysing population histories. I wrap up the introduction by setting out the research questions I sought to answer in this chapter.

3.3.1 Human migrations in the Americas

The Americas were the last of the inhabited continents to be populated. The most recent findings suggest a four wave population model of migrations from Asia (Skoglund and Reich, 2016). Over 15,000 years ago, an ice free corridor opened up between the Cordilleran and Laurentide ice sheets and a substructured population entered the Americas from Beringia via a Pacific coastal route with the majority of modern Native American groups tracing their ancestry to this 'First American' population. This population seems to have spread rapidly and colonised most of both North and South America. Archaeological and genetic evidence suggest that people colonised Arctic regions of the Americas at least twice over the last 6,000 years during two major post-glacial migrations, staring with the Paleo-eskimo migration, reaching as far as Greenland ~4,500 BP. The Paleo-eskimo population was almost entirely replaced by an expansion of the direct ancestors of modern day Inuit populations (Raghavan et al., 2014), the Thule people ~1,000 years ago. Entry into the Americas is associated with a genetic bottleneck, such that modern Native American populations have the lowest levels of genetic diversity of any continental group (Wang et al., 2007).

3.3.2 Dogs in the Americas and Arctic

The earliest confirmed evidence of dogs in North America appears roughly 4,500 years after the first evidence of human activity on the continent (Goebel et al., 2008; Jakobsson et al., 2017). In Central and South America, the earliest dog remains appear in Mexico at around 3,200 BC (Mitchell, 2015) and in Ecuador between 3,000 and 2,500 BC (Stahl, 1985), thousands of years after the first evidence of humans ~14 thousand years ago (kya) (Dillehay and Collins, 1988). Previous studies have suggested that early dogs present in the Americas before European contact (pre-contact dogs) were transported by humans migrating from Asia (Leonard et al., 2002; van Asch et al., 2013) while others suggest these dogs were independently derived *in situ* from American wolves (Koop et al., 2000; Witt et al., 2015). Evidence suggests that dogs played a number of functional roles in early American cultures in fending off predators and hauling goods and people, in addition to being a source of fur and food. In fact, prior to the recent direct radiocarbon dating of the oldest dog remains in the Americas (Perri et al., 2018), the earliest evidence for dogs came from occipital bones (9,260 cal BP) preserved in a human coprolite (Tito et al., 2011).

By the time of European contact, dogs were widespread in the Americas and early European settlers described indigenous dogs as diverse in appearance and pervasive throughout the continent (Ubelaker and Sturtevant, 2007). Studies of the mitochondrial DNA control region suggest that the pre-contact American dog population was largely replaced following the arrival of European dogs after colonisation and the introduction of Eurasian Arctic dogs, such as Siberian huskies, during the 19th century Klondike Gold Rush (Brown et al., 2015; Castroviejo-Fisher et al., 2011; Witt et al., 2015) but the fate of these pre-contact dogs remains largely unknown. It has been suggested that some modern dogs in the Americas may retain a degree of ancestry from the pre-contact population including the Carolina dog, chihuahua, and Mexican and Peruvian hairless dogs (Shannon et al., 2015; van Asch et al., 2013). This suggestion remained controversial, however, since the nuclear ancestry of pre-contact American dogs had not been established.

Arctic dogs represent a highly genetically divergent, basal lineage of modern dogs (Larson et al., 2012; Vonholdt et al., 2010). It is suggested that they are likely to have undergone recent admixture with wild canids, given their close proximity to extensive wolf and coyote populations (Vonholdt et al., 2010; Wang et al., 2016). Modern

high-latitude dog breeds, like the Greenland sledge dog and Siberian husky, show evidence of introgression from a 35,000 year old wolf from the Taimyr Peninsula in Northern Siberia (Skoglund et al., 2015). Malamutes, Greenland dogs and huskies also show strong signals of admixture with Eurasian dogs, in particular with East Asian lineages (Frantz et al., 2016; Wang et al., 2016). It has previously been suggested that this is due to several waves of dogs migrations to the Americas alongside human colonisers at different time periods (Wang et al., 2016).

3.3.3 Inferring population histories from genetic data

A number of different tests are applied to genetic data sets in this chapter in order to analyse the population history of the CTVT founder alongside early dogs in the Americas, and so a brief overview of these methods is in order. The majority of the methods used are based on allele frequencies at variant sites and the general assumption that similar allele frequencies indicate that individuals or populations have a shared ancestry. For example, principal component analysis (PCA), based on this assumption, allows unsupervised clustering of individuals based on their ancestry (Patterson et al., 2006). Model-based clustering methods implemented in STRUCTURE (Hubisz et al., 2009) and ADMIXTURE (Alexander et al., 2009) are commonly used approaches that assign individual ancestry based on a specified number of ancestral components.

Allele frequency differentiation statistics (f-statistics; Green et al., 2010), comprising the f_2, f_3 and f_4 statistics, and a statistic very similar to the four-population test known as the *D*-statistic (Patterson et al., 2012), are a standard means of describing patterns of allele frequency correlations across populations. It is worth nothing that at various points throughout this chapter, the *D*-statistic is also referred to as the ABBA BABA test. Here 'A' is the ancestral allele and 'B' is the derived allele; 'ABBA' and 'BABA' refer to genotype patterns. The three and four-population tests can be used as formal tests of admixture and, in the case of the f_4 or *D*-statistic tests, provide information about the directionality of gene flow. These methods can be applied to most genetic data, as they rely only on information about the allele frequencies observed in populations, without any marked constraints on the density or number of markers, excepting that markers are not in strong linkage disequilibrium. Various other methods, such as PCAdmix (Brisbin et al., 2012) and HAPMIX can additionally

make use of genetic linkage information to identify the ancestry of chromosomal segments from individuals in admixed populations. However, these haplotype based methods require phased data and, in order to exploit linkage information, haplotype-based inference methods also require a high density of variants, such as that obtained by whole-genome sequencing or high-density genotyping arrays.

3.3.4 Summary

The primary aim of this chapter was an understanding of the dog population to which the CTVT founder belonged. To begin with, the study was motivated by the following questions:

(i) What is the likely geographic origin of CTVT?

(ii) What is the temporal origin of CTVT?

(iii) Which ancient or modern dog population is the CTVT founder most closely related to?

The discovery of a genetic affinity between the CTVT founder and native American dogs lead to more detailed questions regarding the inceptive host population:

(i) What is the origin of pre-contact dogs?

(ii) What was the fate of pre-contact dogs?

(iii) What is the genetic contribution of pre-contact dogs to modern dog populations?

(iv) How are pre-contact dogs related to other ancient and modern dog populations?

(v) Previous studies have divided modern dog populations into two "core" groups of East Asian and West Eurasian dog groups; where do pre-contact dogs fit in?

(vi) What was the interaction between pre-contact dogs and other canid populations in the Americas?

A large part of this chapter is centred on the above questions related to the enigmatic pre-contact American dog population.

3.4 Methods

3.4.1 Archaeological background

This section presents site descriptions for all ancient dog samples sequenced in this study. Archaeological site descriptions were provided by Angela Perri and Carly Ameen and I have adapted these for this chapter. Morphological descriptions of canid remains as well as the historical details provided give a necessary and relevant context for interpretation of the genetic analyses that follow. Study sample numbers appear in parentheses. The nomenclature 'Canis spp.' is used to refer to two or more unspecified species of the genus Canis.

3.4.1.1 USA

Cox, Alabama (Cox6)
The Cox mound site is located in Jackson County, Alabama and dates to the Middle Woodland period (AD 1-500) (Moore, 1915). Multiple dogs have been recovered from the site, one of which is included in this analysis.

Flint River, Alabama (FR11)
The Flint River site was a village with a large circular shell mound (2 m tall, 15 m in diameter) located on the Flint River near Courtland, Alabama (Moore, 1915). Over 200 human burials have been recovered from the site, as well as 19 dog burials (Webb and DeJarnette, 1948a). Dogs were interred throughout the period of occupation, and did not share graves with humans. The dog analysed in this study may be Late Archaic or Mississippian (Warren, 2004).

Little Bear, Alabama (LB2)
The Little Bear Creek site is an Archaic shell mound located at the mouth of the Little Bear Creek, in Colbert County, Alabama (Webb and DeJarnette, 1948b). Dog remains recovered from this site were all dated to the Late Archaic period, and were deliberately buried in the mound, similar to human interments, however dogs and humans were not found buried together at this site (Barkalow, 1972).

Perry, Alabama (P35, P59)
The Perry site is a large shell mound located in northern Alabama, on the Tennessee

River (Webb and DeJarnette, 1948c). The shell mound contains hundreds of human burials that have been dated to the middle Archaic to Mississippian periods, based on the artefacts present with the burials, as well as a large assemblage of stone tools (Futato, 2002). Over 100 dogs have also been buried in the shell mound, some of them with humans. The age of these dog burials is largely unknown, as the majority of the dogs were buried without associated grave goods.

Channel Islands, California (CINHA, CINH7, CIAS, CICVD, CAW2, CAO1)
The Channel Islands are located in the Pacific Ocean, off the coast of southern California, and have been occupied for 13,000 years (Rick et al., 2005). Dogs have lived on the islands for at least 6,000 years, and were almost certainly introduced by humans (Rick et al., 2008). They were likely not used as hunters, but may have been an occasional food source, and were found on all of the Channel Islands. Six individuals from two islands, as well as one mainland archaeological site, were included in this study. Two dogs were recovered from Santa Cruz Island (CAW2, CAO1), one from the Orizaba Cove area that dates to the Late Holocene, and the other from Willows Canyon. Two dogs are from the North Head site on San Nicolas Island (CINH4, CINH7), which was occupied during the Terminal Early Period (5,000 BP) and the Middle Period (2,000 BP). One dog derives from Santa Rosa Island (CICVD), at the Canada Verde site, which dates to the terminal Early Period (4,000 BP). The dog from the mainland site (CIAS) was recovered from the coastal Chumash town of Syuxtun (now a part of modern-day Santa Barbara).

Grass Mesa, Colorado (5MT316)
Grass Mesa Village was a large village located east of the Dolores River in Colorado (Lipe et al., 1988). It was occupied from AD 700 until roughly AD 900, and has one of the highest artefact densities of any archaeological site in the Mesa Verde region. The majority of faunal remains from the site were of multiple species of deer and rabbits, although there were also small numbers of dog and unidentifiable Canis spp. bones.

McPhee Pueblo, Colorado (5MT520)
McPhee Pueblo was part of McPhee Village, which is located 6 km northwest of Dolores, Colorado (Kane and Robinson, 1988). It was occupied for 200 years during the Pueblo I and II periods (roughly AD 800-980). Over 200 bones have been recovered

from the site that are identified as Canis spp., including coyotes, wolves, and dogs.

Yellow Jacket Pueblo, Colorado (5MT501)
Yellow Jacket Pueblo was a large village in the Central Mesa Verde located near modern-day Cortez, Colorado. Yellow Jacket was occupied during the Late Pueblo II through the Pueblo III periods (~AD 1050-1260) (Kuckelman, 2013). A small number of canid remains were identified at the site, including one bone that was specifically identified as belonging to a dog. Some of the canid bones show evidence of burning or cut marks, suggesting that these dogs were used as food.

Baum, Ohio (AL2748)
The Baum site is a prehistoric village settlement in Paint Creek River Valley, Ohio. This is the type site for the Baum Phase of the Fort Ancient culture, dating between AD 950-1200 (Hart and Rieth, 2002; Trigger and Sturtevant, 1978). Abundant, well preserved faunal remains were found at the site within refuse pits, and remains of dogs were found ubiquitously throughout the village. These remains were described as belonging to Indian dogs of a size and proportion similar to the bull terrier (Mills, 1906). A total of 50 dog remains, including 7 crania, were collected. All crania collected were broken; it has been suggested that this was done in order to remove the brain (Mills, 1906). Comparisons with other prehistoric dogs suggest these were of the general Indian dog type found at contemporary sites in the region, as well as at prehistoric sites in Texas (Mills, 1906).

While individual dog burials and associated ritual activity are known from contemporary sites (Morey, 2006; Nolan and Sciulli, 2014), the evidence from Baum suggests these dogs played a more domestic role. The breakage of dog bones, similar to other food animals, indicates that at least on some occasions dogs were used as food. Gnaw marks from dogs on other animal bones found throughout the middens suggests that dogs had access to food waste. Unlike earlier Archaic period sites where dogs were regularly afforded special treatment indicated by their careful burial, later prehistoric dogs from Baum appear to have been a utilitarian presence at the site, likely similar to modern feral or village dogs, with deposition predominantly occurring in refuse pits.

Reinhardt, Ohio (AL2772)
The Reinhardt site is a Fort Ancient village site in the Scioto Valley, Ohio (Nolan,

2009, 2011). Associated AMS dates from wood, charcoal and a single human burial support the dating of the site to within the Middle Fort Ancient period (AD 1200-1400) (Nolan, 2011), though bifaces from at least four earlier time periods have been recovered at the site.

Excavations in 2008 revealed a small number of canid remains, including two individual dog burials (Nolan, 2009; Nolan and Sciulli, 2014). The specimen in this study is from a fully articulated, adult, male dog with heavily worn teeth, that was found buried under a layer of sand with a turkey bone awl (Nolan, 2009). Significant pathologies to the dog were observed, including vertebral pathologies representing either severe arthritis, or a healed infection, and an abscess in the right M2 (Nolan and Sciulli, 2014). Analysis of dog burials at the Reinhardt site showed significant variation between dogs (Nolan and Sciulli, 2014). Comparisons with other dogs from the region indicate little change in morphology between the Archaic and later prehistoric periods, despite intense changes in subsistence patterns and lifestyle (Nolan and Cook, 2010; Nolan and Sciulli, 2014).

Dogs in this region were known to serve a variety of purposes, such as for transportation, hunting, companionship and as a food source. Dogs also participated in ceremonial and ritual activity, and dog sacrifices were not uncommon (Cook, 2012; Kerber, 1997; Morey, 2006). The individual burial of dogs at Reinhardt suggests a level of affinity remained between the Fort Ancient people and at least some of their dogs, though isolated dog remains found throughout the site suggests that this intimacy was not afforded to all dogs equally.

Scioto Caverns, Ohio (OSU611, OSU622, OSU624, OSU626, OSU628, OSU634, OSU638)
The Scioto Caverns sites in Ohio are a series of three limestone caverns located near the Scioto River and Wright-Holder earthworks complex (Potter and Baby, 1964). The bones of 25 dogs have been recovered from this deposit, including 5 nearly-complete skulls and 11 mandibles. The dogs are likely from the Middle Woodland or Hopewellian period, based on the nature of their burial. Based on their size, the dogs appear similar to Archaic dogs identified from Kentucky and Alabama, and are smaller than Woodland period dogs from the Midwest.

Apple Creek, Illinois (ISM070)

The Apple Creek site is a Woodland village on the north bank of Apple Creek, four miles north of Eldred, Illinois (Parmalee et al., 1972). Seven dog burials have been recovered from the site, as well as some additional isolated elements, which show evidence of consumption. The dog burials likely date from the Middle or Late Woodland periods (2,000-1,000 BP). The dogs are all terrier-sized, which is smaller on average than most other Woodland period dogs. The dog used in this analysis was a small, mature, male dog with more tooth wear on the right side of the maxilla and mandible than the left.

Koster, Illinois (AL2135)

The Koster site is a complex, highly-stratified site located in a tributary valley of the lower Illinois River in west-central Illinois (Butzer, 1978; Chapman et al., 1981; Hajic, 1990; Houart, 1971). Cultural deposits dating from the early Archaic through Mississippian periods were buried to depths of over 10m, providing a continuous record of Holocene human occupation in the region (Brown and Vierra, 1983).

Excavators uncovered the burial of four dogs, found as complete, articulated skeletons in shallow pits (Hill, 1972; Morey and Wiant, 1992). At the time of excavation the dogs were dated based on associated material to around 8,500 years ago. Recent direct dating of two of the Koster dog burials dated them to between 10,130-9,680 cal BP respectively (Perri et al., 2018), making them among the earliest dated material from Koster, the earliest identified dogs in the Americas and the earliest dog burials in the world (Fig. 3.1).

The specimen included in this study (AL2135) was found in a shallow, basin-shaped pit with a metate and mano placed near its cranium, though it is not clear whether this is associated with the burial. The skeleton is complete and the animal was buried lying on its side, with no evidence of intentional cut marks or other trauma. Given the presence of preserved bacula in the other dogs, the missing baculum here suggests the dog was a female, and that it was an adult (Morey and Wiant, 1992). Morphological comparisons of Koster canids with modern and ancient wolves and coyotes suggested them to be domesticated dogs, based on their small size (Morey and Wiant, 1992). In particular, a wide palate and cranial vault, distinguishing this specimen from coyotes of a similar size, was noted.

Janey B. Goode, Illinois (JBG1m, JBG5, JBG11, JBG12, JBG13, JBG17, JBG19, JBG21, JBG24, JBG26, JBG32, JBG35, JBG37, JBG41, JBG42, JBG43, JBG45, JBG48, JBG50)
The Janey B. Goode site is a settlement near Brooklyn, Illinois, that was occupied from AD 650-1400. Janey B. Goode has one of the largest numbers of dog burials of any archaeological site: dogs have been recovered from 102 different features (Borgic and Galloy, 2004). The majority of these dogs have been recovered from the Terminal Late Woodland component (AD 650-900). Many of these dogs were recovered as complete skeletons, and were found in storage pits or buried underneath settlement structures. Cut marks on a Mississippian dog indicate that dogs may have been consumed during this period; dogs from the Late Woodland and Terminal Late Woodland periods show no evidence of consumption, but instead have vertebral fractures, which may be suggestive of their use as pack animals prior to the Mississippian period.

Modoc, Illinois (AL2706)
The Modoc Rock Shelter, located 2 miles south east of Prairie du Rocher, Illinois, is one of the earliest sites of human occupation in Illinois (Fowler, 1959). The lowest strata has been dated to 11,000-9,000 BP, and the site was occupied for 6,000 years. Post moulds have been identified starting at 8,000 BP, along with six human burials, but there is no trace of human occupation after 4,000 BP. Two complete dog burials were recovered from the same strata as the human burials, one of which is included in this study (Fowler, 1959).

Angel Mounds, Indiana (AM310A, AM310B, AM310C, AM474)
Angel Mounds is located along the Ohio River east of Evansville, Indiana. Site construction began just before AD 1050 and continued through AD 1450 with the primary occupation occurring during the Late Mississippian period between AD 1350-1450 (Monaghan and Peebles, 2010).Several dog burials and dog skulls are recorded at Angel Mounds. Of the four dog burials listed in the Angel Mounds catalogue, only one contains an articulated dog skeleton that can be considered as an intentional and undisturbed dog burial (AM310A). There are also the remains of two dogs (which are mixed in the same context) recorded as "dog burial", AM310B and AM310C, including a large dog's limb bones and postcranial elements as well as a smaller dog's cranial and mandibular fragments (partially burned). These second and third "burials" consist of several disarticulated bone fragments representing incomplete

remains, and do not have any hand-drawn or photographed *in situ* representations associated with them. AM310B and AM310C come from the same excavation block as the articulated dog and AM474. Materials associated with AM474 include the cranium of an average-sized Mississippian dog, without mandibles or postcranial elements.

Simonsen Bison Kill, Iowa (ISM172)
The Simonsen site is located in northwest Iowa, near Quimby, and its usage dates to the Paleoindian and Archaic periods (Frankforter and Agogino, 1960). It has been suggested to be a bison kill site, given that the majority of remains at the site are of bison (Agogino and Frankforter, 1960). The highest zone at the site contained charred wood and ash suggestive of a firepit, as well as the ramus of a large dog, which was included in this analysis.

Weyanoke Old Town, Virginia (AL3226, AL3223)
The Weyanoke Old Town site, or Hatch site, is located on a tributary of the James River on the coastal plain southeast of Richmond, Virginia and covers over 4,329 sq. metres (Buck, 2000). Based on associated artefacts, occupation at the site spans from at least the early Archaic through the early English Colonial period, when the Virginia Algonquians occupied the region (Gregory, 1980). The site was home to a Weyanoke (Weanoc tribe) village dating from the prehistoric through the early Colonial period (McCary, 1995).

Over 112 domestic dogs have been recovered from the site, one of the largest discoveries of domestic dogs in the Americas (Buck, 2000; Kerber, 1997). The dog was the only domesticated animal of the Virginia Algonquians (Robinson and Rountree, 1991). There is no evidence that dogs at Weyanoke were used for food, as all dogs were recovered in isolated and associated burials, articulated with no evidence of cut marks or burning (Buck, 2000). Dogs of the Virginia Algonquians are documented partaking in small and large game hunting and as protection from predators like wolves and bears (Trigger and Sturtevant, 1978). Like other regions of the Midcontinent and Eastern Woodlands (Strong, 1985), dogs played an important role in burial with humans at Weyanoke. It is unclear whether dogs at Weyanoke were viewed as companions, ritual offerings, or both, but dog sacrifices were not uncommon among contact-period tribes (Wallis, 1955).

Descriptions of the Virginian native dog often mention a wolf-like (Harriot, 1588; Smith, 1632) or fox-like (Brereton and Haies, 1602) appearance with an inability to bark and a propensity for howling. In an analysis of the Weyanoke Old Town skeletal material, Blick (Blick, 1988) noted the dogs there are clearly domesticated and not wolf-like in their cranial and postcranial skeletal morphology. He later proposed that admixture between local wolves and aboriginal dogs may explain the reported "wolf-like" behaviour and appearance of Algonquian dogs (Buck, 2000). They stood an average of 42 cm high and weighed approximately 10kg, similar to a medium-sized Algonquian native village dog depicted in a 1585 painting by John White (Hulton, 1984). This painting, the earliest depiction of a Native American dog by a European, shows a knee-height, medium-sized yellow-brown dog with pricked ears and a long tail, similar to many modern village dogs or dingoes. In contrast to this, in 1602, John Brereton described dogs in coastal Massachusetts as "fox-like, black, and sharp-nosed" (Brereton and Haies, 1602). The Weyanoke canids are associated with Late Woodland period artefacts, confirming their native ancestry and lack of inbreeding with later introduced European breeds. This is supported by direct radiocarbon dating of one of the dogs used for DNA analysis (AL3223) to 985-935 cal BP.

Uyak, Alaska (AL3198)

The Uyak Site is a substantial prehistoric midden covering hundreds of acres and located on the western side of Kodiak Island, Alaska. The site was excavated between 1933-1936 for the United States National Museum (now the National Museum of Natural History), and consisted of house structures and hearths, stone and organic artefacts, human remains, and a large faunal assemblage (Hrdlička, 1944). Occupation of the site is suggested to date from 2,000 BP to Russian contact in the mid-18th century (Heizer, 1956; Hrdlička, 1944). Due to the long occupation history of the site, and without direct dating, the cultural context of the study specimen can only be inferred to span both the earlier Kachemak phase (4,000-900 cal BP) and the later Koniag phase (900-200 cal BP), as it is clear dogs were present throughout the occupation (Allen, 1939; Hrdlička, 1944). Both cultures were maritime hunter-gatherers (Hrdlička, 1944; Scott, 1992).

Dogs were ubiquitous throughout the site and hundreds of domestic dog remains have been collected. Early metric analysis identified two types of dogs, classified as

a "small" and a "large" type (Allen, 1939). While it was originally suggested that the small type was restricted to earliest deposits (Hrdlička, 1944) and later replaced by the large type, other studies have noted size variability among contemporary prehistoric canids throughout the Kodiak archipelago (Clark and Milan, 1974), and more recent work suggests the variation could be a result of sexual dimorphism (West and Jarvis, 2015).

Excavation of the specimens was poorly recorded, so it is unknown whether the dog remains represent intentional burials, midden/refuse deposits or some combination of deposition types. This makes an interpretation of the role of dogs at the site difficult, though cut marks suggesting butchery were recorded on a portion of both mandibular and cranial elements (West and Jarvis, 2015). As the region's sole domesticate, it is likely that dogs at Uyak fulfilled several of roles simultaneously including as a food source and additionally serve a variety of other functions in the Arctic, including assisting in prey acquisition, transportation, sanitation through food waste disposal, and even warmth (Groves, 1999).

3.4.1.2 Canada

Prince Rupert Harbour, British Columbia (PRD1, PRD9, PRD10, PRW89)
Prince Rupert Harbour is on the northern Northwest Coast of North America, just south of the Alaskan boundary. Two sites, GbTo-13 and GbTo-54, had large-scale excavations in 2012-2013 (Eldridge et al., 2014). Components at the site date from 900 BC (radiocarbon calibrated) to about 250 BC; the main occupation was between 250 BC and AD 800, followed by slightly less intensive occupation from about AD 900-1300. Some late European contact activity also occurred, but there were no European trade goods at the site, suggesting no admixture with European dogs likely occurred in the population of dogs at this site. GbTo-13 had a smaller occupation limited to AD 1,000–1,300.

An unprecedented number of bones from mountain goat (Eldridge et al., 2014), grizzly and black bear, sea lion, northern fur seal, wolf, and other species otherwise rare in Northwest Coast animal assemblages were found at both sites. A single purposeful dog burial was found at GbTo-54, outside a house identified as a chief's residence based on rank-linked artefact and faunal distribution. Some 300 dog bones

and 21 wolf bones were recovered from GbTo-54, along with 25 dog bones from GbTo-13. Remains indicated that dogs were smaller than the wolves in the area, and so the two species could be distinguished. Three dogs and a wolf were included in this analysis.

Port au Choix, Newfoundland (AL3194)
The Port au Choix site is located on Newfoundland's northwest coast on the Port au Choix peninsula. The area encompasses a number of well-preserved localities, including a Maritime Archaic burial ground (Port au Choix-3) with over 100 preserved burials (Port au Choix-3, Locus II), excavated from 1967-1969 by Memorial University of Newfoundland (Tuck, 1970, 1971, 1976). The burial ground at Port au Choix is thought to date to approximately 4,400-3,300 years ago (Renouf, 1993).The remains of four dogs were recovered from the Port au Choix-3 burial ground including two complete and articulated associated burials accompanying a human burial. All the dogs appeared to be of the Large or Common Indian dog size, though there was some variation between them, within the range of sexual dimorphism (Tuck, 1976).

The dog analysed in this chapter (AL3194) is an older male, likely weighing between 20-25 kilograms, killed by a blow to the head, and included, along with another male dog, in a multi-human burial (Tuck, 1976). Direct radiocarbon dating of the dog resulted in a date of 4,402-3,912 cal BP, indicating it comes from the earliest use of the burial ground within the Maritime Archaic period. These dogs were likely used as companions, hunting aids, and occasionally as travois dogs due to their well-developed muscle attachments (Tuck, 1976). Comparisons to Eskimo sledge dogs were unfavourable, but the Port au Choix specimens were similar to other large breed native dogs seen at the prehistoric Frontenac Island site in New York (Smith, 1947). Early descriptions of large breed native dogs described them as slender and wolf-like, with large erect ears and a long, pointed snout (Allen, 1920).

3.4.1.3 Central America

Mayapan, Mexico (May2, May3, May4, May10)
Mayapan, located in Yucatan, Mexico, was the largest Mayan city during the Post-classic period, and was occupied from its founding in 900 BP until its abandonment in 450 BP (Lope et al., 2006). It consisted of a monumental centre containing temples

and ritual buildings, surrounded by residential areas (Pollock, 1962). Over 2,000 dog bone fragments have been identified from the site (Masson and Peraza Lope, 2008). Dog remains were highly concentrated in the Templo Redondo Group, located in the main plaza of the monumental centre (Masson and Peraza Lope, 2013). There is evidence that young dogs were consumed regularly, and that older dogs were often used in ritual contexts.

3.4.1.4 Russia

Aachim Lighthouse, Russia (CGG10, CGG11)
The Aachim Lighthouse site is located in the Aachim Peninsula on the East Siberian Sea coast. Mandibles of two dogs included in this study were radiocarbon dated to 1,760 ± 40 BP and 1,740 ± 40 (Lee et al., 2015).

Zhokhov, Siberia (CGG1, CGG2, CGG3, CGG4, CGG5, CGG6, CGG7, CGG8, CGG9)
Zhokhov Island is located off of the northeastern coast of Siberia (Pitulko and Kasparov, 2017)). Human occupation of Zhokhov began in the Late Pleistocene period, with the earliest evidence for humans at the site dating to 9,000 years ago (Pitulko and Kasparov, 2017). Small blades and bone and ivory points have also been recovered from the site, as well as a fairly sophisticated sledge runner for the time period. Two fragmented dog mandibles, as well as a small number of postcranial bones, have been recovered from the site, and date to the earliest period of human occupation. They are smaller than wolf mandibles, and are similar in size to other ancient Arctic dog remains that have been recovered.

Previous mitochondrial DNA sequencing demonstrated that Zhokhov dogs belong to dog Haplogroup A, and are genetically indistinguishable from modern domestic dogs in the hypervariable region of the mitochondrial genome (Lee et al., 2015).

3.5 Ancient DNA

3.5.1 DNA extraction and sequencing

Ancient DNA (aDNA) extractions and library preparations were carried out at the University of Oxford under the supervision of Laurent Frantz and Greger Larson, at University of Illinois at Urbana-Champaign under the supervision of Ripan Malhi,

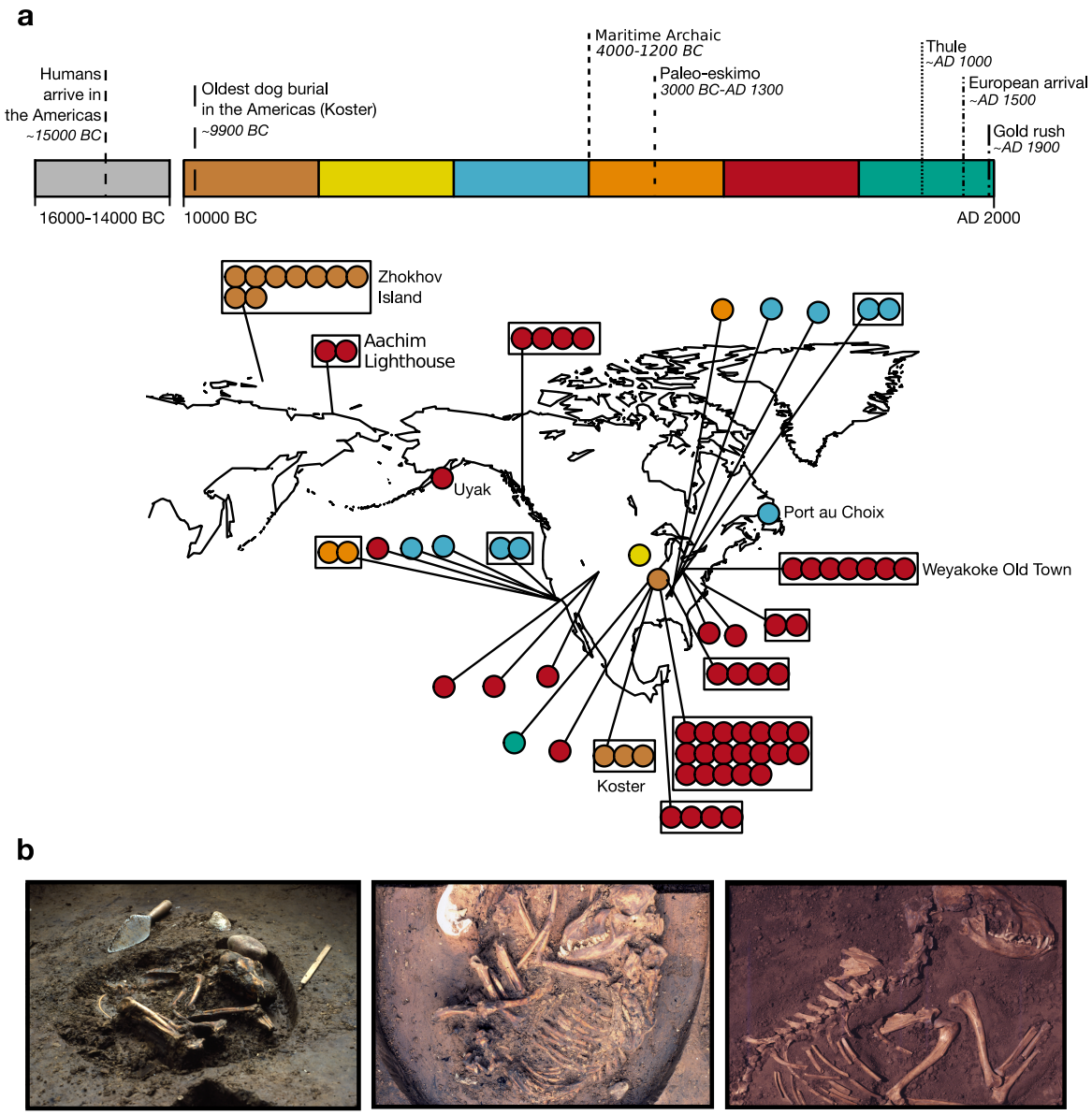


Fig. 3.1. a. A map depicting the locations and ages of the archaeological remains analysed in this study. Each dot represents a single sample, and multiple samples per archaeological site are grouped in boxes. Key sites mentioned are labelled. BP, before the present. Figure adapted from Ní Leathlobhair et al. (2018). **b.** *In situ* photographs of dog burials at Koster site, Illinois. Images provided by Del Baston, courtesy of the Center for American Archaeology.

Table 3.1 Summary information about ancient dog samples analysed in this chapter including age (based on radiocarbon dating or stratigraphic context). ND (not determined) appears for samples where dating was not performed, or for which osteological sex was undetermined.

Sample ID	Age (radiocarbon)	Age (context)	Site	Latitude	Longitude	Sex
AL2135	10110-9680 cal BP	Archaic, Transition	Koster, Illinois, USA	39.209	-90.549	F
AL2748	ND	1000-700 BP	Baum Village, Ohio, USA	39.282	-83.157	M
AL2772	ND	750-550 BP	Reinhardt, Ohio, USA	39.779	-83.004	M
AL3194	4402-3912 cal BP	Maritime Archaic	Port au Choix, Canada	50.703	-57.352	M
AL3198	ND	2000-200 BP	Uyak, Alaska, USA	57.519	-154.017	M
AL3223	985-935 cal BP	AD 680-1430	Weyanoke Old Town, Virginia, USA	37.289	-77.066	F
AL3226	ND	750-1450	Weyanoke Old Town, Virginia, USA	37.290	-77.303	F
5MT316	ND	1100-1400 BP	Grass Mesa, Colorado, USA	39.469	-107.771	ND
5MT501	ND	800-1000 BP	Yellow Jacket Pueblo, Colorado, USA	37.56	-108.711	ND
5MT520	ND	1100-1300 BP	McPhee Pueblo, Colorado, USA	37.47	-108.5	ND
AM310A	ND	1000 BP	Angel Mounds, Indiana, USA	37.943	-87.458	ND
AM310B	ND	1000 BP	Angel Mounds, Indiana, USA	37.943	-87.458	ND
AM310C	ND	1000 BP	Angel Mounds, Indiana, USA	37.943	-87.458	ND
AM474	ND	1000 BP	Angel Mounds, Indiana, USA	37.943	-87.458	ND
CAO1	ND	2000-6000 BP	Channel Islands, California, USA	34.045	-119.723	ND
CAW2	ND	2000-6000 BP	Channel Islands, California, USA	33.961	-119.755	ND
CGG1	ND	7900-9000 BP	Zhokhov site, Russia	76.141	152.733	ND
CGG10	1720-1780 cal BP	ND	Aachim Lighthouse, Russia	69.881	173.4682	ND
CGG11	1700-1740 cal BP	ND	Aachim Lighthouse, Russia	69.881	173.468	ND
CGG2	8910-8960 cal BP	7900-9000 BP	Zhokhov site, Russia	76.141	152.733	ND
CGG3	ND	7900-9000 BP	Zhokhov site, Russia	76.141	152.733	ND
CGG4	ND	7900-9000 BP	Zhokhov site, Russia	76.141	152.733	ND
CGG5	ND	7900-9000 BP	Zhokhov site, Russia	76.141	152.733	ND
CGG6	ND	7900-9000 BP	Zhokhov site, Russia	76.141	152.733	ND
CGG7	ND	7900-9000 BP	Zhokhov site, Russia	76.141	152.733	ND
CGG8	ND	7900-9000 BP	Zhokhov site, Russia	76.141	152.733	ND
CGG9	ND	7900-9000 BP	Zhokhov site, Russia	76.141	152.733	ND
CIAS	ND	2000-6000 BP	Syuxtun, California, USA	34.410	-119.691	ND
CICVD	ND	4000 BP	Canada Verde Site, Santa Rosa Island, California, USA	34.024	-120.132	ND
CINH7	ND	5000 BP or 2000 BP	North Head site, San Nicholas Island, California, USA	33.27	-119.566	ND
CINHA	ND	5000 BP or 2000 BP	North Head site, San Nicholas Island, California, USA	33.27	-119.566	ND
CISG	ND	2000 BP or 700 BP	Channel Islands, California, USA	34.005	-120.181	ND
Cox6	ND	1500-3000 BP	Cox Site, Alabama, USA	34.824	-86.011	ND
FR11	ND	3000-7000 BP	Flint River Site, Alabama, USA	34.977	-86.539	ND
ISM070	ND	2500-1000 BP	Apple Creek, Illinois, USA	40.145	-89.172	ND
AL2706	8560-8210 cal BP	ND	Modoc, Missouri, USA	38.063	-90.064	ND
AL2135	10110-9680 cal BP	Archaic, Transition	Koster, Illinois, USA	39.209	-90.549	ND
AL2704	10130-9680 cal BP	Archaic, Transition	Koster, Illinois, USA	39.209	-90.549	ND
JBG11	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG12	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG13	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG17	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG19	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG1m	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG21	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG24	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG26	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG32	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG35	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG37	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG41	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG42	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG43	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG45	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG48	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG5	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
JBG50	ND	1000-1400 BP	Janey B. Goode, Illinois, USA	38.658	-90.162	ND
LB2	ND	3000-7000 BP	Little Bear, Alabama	34.307	-87.665	ND

Table 3.1 continued

Sample ID	Age (radiocarbon)	Age (context)	Site	Latitude	Longitude	Sex
May10	ND	1000 BP	Mayapan, Mexico	20.461	-89.217	ND
May2	ND	1000 BP	Mayapan, Mexico	20.461	-89.217	ND
May3	ND	1000 BP	Mayapan, Mexico	20.461	-89.217	ND
May4	ND	1000 BP	Mayapan, Mexico	20.461	-89.217	ND
OSU611	ND	2200-1600 BP	Scioto Caverns, Ohio	40.113	-83.107	ND
OSU622	ND	2200-1600 BP	Scioto Caverns, Ohio	40.113	-83.107	ND
OSU624	ND	2200-1600 BP	Scioto Caverns, Ohio	40.113	-83.107	ND
OSU626	ND	2200-1600 BP	Scioto Caverns, Ohio	40.113	-83.107	ND
OSU628	ND	2200-1600 BP	Scioto Caverns, Ohio	40.113	-83.107	ND
OSU634	ND	2200-1600 BP	Scioto Caverns, Ohio, USA	40.113	-83.107	ND
OSU638	ND	2200-1600 BP	Scioto Caverns, Ohio	40.113	-83.107	ND
P35	ND	3000-7000 BP	Perry, Alabama	34.915	-87.684	ND
P59	ND	3000-7000 BP	Perry, Alabama	34.915	-87.684	ND
PRD1	ND	1500 BP	Prince Rupert Harbour, Canada	54.305	-130.343	ND
PRD10	ND	1500 BP	Prince Rupert Harbour, Canada	54.305	-130.343	ND
PRD9	ND	1500 BP	Prince Rupert Harbour, Canada	54.305	-130.343	ND

and at University of Copenhagen under the supervision of Tom Gilbert. DNA was extracted from canid teeth or bone samples in a dedicated ancient DNA laboratories using the appropriate sterile techniques and equipment. Most of the DNA extractions were performed at the University of Illinois at Urbana Champaign, at the Carl R. Woese Institute for Genomic Biology, with methods described in Witt et al. (2015)), but a subset of extractions was performed at the Centre for GeoGenetics at the University of Copenhagen, with methods described in Allentoft et al. (2015) and at the University of Oxford following the protocol described in Dabney et al. (2013). Ancient DNA laboratory work was performed by Anna Linderholm, James Haile, Ophelie Lebrasseur, Evangelos Dimopoulos (University of Oxford), Kelsey Witt (University of Illinois at Urbana-Champaign), Selina Brace (Natural History Museum, London), Jacob Enk, Alison Devault, Jean-Marie Rouillard (Arbor Biosciences) and Mikkel-Holger Sinding (University of Copenhagen).

Libraries were sequenced on an Illumina HiSeq 2500 system (single-end 80 bp) at the Danish National High-Throughput Sequencing Centre, on an Illumina NextSeq 500 system (single-end 80 bp) at the Natural History Museum (London) and on an Illumina HiSeq 2500 system (80 or 100 bp) at the Roy J. Carver Biotechnology Center at the University of Illinois.

Table 3.2 Sequencing statistics for ancient dog samples included in mitochondrial DNA analyses.

Sample ID	Extraction lab	Number of reads	Endogenous content	mtDNA coverage
AL2772	Oxford	13783057	4.596462164	5.058528128
AL3194	Oxford	265763856	38.33614756	116.2334549
AL3223	Oxford	125598030	16.89122194	71.07419143
5MT316	Illinois	19512	56.02706027	39.31751061
5MT501	Illinois	70113	60.39536177	76.37053865
5MT520	Illinois	4794	25.40675845	5.886829677
AM310A	Illinois	32097	73.98510764	102.7144736
AM310B	Illinois	22394	64.75395195	61.83386142
AM310C	Illinois	25562	48.4547375	47.30860286
AM474	Illinois	23613	53.65264896	47.62306451
CAO1	Illinois	40840	73.14887365	110.2931189
CAW2	Illinois	25137	65.05151768	62.58982483
CGG1	Copenhagen	46062	56.70400764	99.49967119
CGG10	Copenhagen	60830	29.12214368	28.83158965
CGG11	Copenhagen	143208	29.26791799	96.7221259
CGG2	Copenhagen	120267	27.17453666	69.40072936
CGG3	Copenhagen	68821	28.64532628	40.03258205
CGG4	Copenhagen	233112	26.71634236	85.20930233
CGG5	Copenhagen	378999	38.50986414	129.2819394
CGG6	Copenhagen	4946724	74.33394303	57.72637054
CGG7	Copenhagen	138160	59.1806601	145.7030549
CGG8	Copenhagen	44351	27.36578657	31.15268727
CGG9	Copenhagen	83220	33.81639029	87.44449094
CIAS	Illinois	16396	59.2461576	33.38847373
CICVD	Illinois	892127	16.49709066	22.52561727
CINH7	Illinois	20560	63.92509728	47.10402344
CINHA	Illinois	20083	68.31150724	46.00454355
CISG	Illinois	36655	73.758	97.65995098
Cox6	Illinois	24459	26.281	16.97871704
FR11	Illinois	17810	13.981	7.0521
ISM070	Illinois	10188	55.477	26.36659293
AL2706	Oxford/Illinois	18608	69.34114359	52.34136426
AL2135	Oxford/Illinois	33094	70.732	71.35296228
AL2704	Oxford/Illinois	31019	67.575	77.18831829
JBG11	Copenhagen	130420	53.779	162.6978538
JBG12	Copenhagen	185720	34.443	153.6255156
JBG13	Copenhagen	114795	45.01241343	157.7362348
JBG17	Copenhagen	41931	79.16338747	165.3036408
JBG19	Copenhagen	83483	44.75282393	129.5464817
JBG1m	Illinois	58193	52.39977317	55.8328451
JBG21	Copenhagen	25539	68.63228787	79.32217373
JBG24	Copenhagen	54470	71.94969708	154.409398
JBG26	Copenhagen	43857	71.20414073	137.4965027
JBG32	Copenhagen	54179	71.94484948	140.0304896
JBG35	Copenhagen	28482	34.8676357	25.7563819
JBG37	Copenhagen	73435	45.53686934	128.6252167
JBG41	Copenhagen	164370	37.65833181	168.990614
JBG42	Copenhagen	152682	60.71639093	158.2733903
JBG43	Copenhagen	670175	64.4883799	193.1227955
JBG45	Copenhagen	73704	56.19098014	154.8202905
JBG48	Copenhagen	65927	51.99690567	149.6821905
JBG5	Copenhagen	51284	54.12214336	122.7431099
JBG50	Copenhagen	31422	41.09222838	50.2458301
LB2	Copenhagen	101269	23.50373757	44.6299994
May10	Illinois	16488	24.83624454	12.72218569
May2	Illinois	31184	56.94907645	61.54397083
May3	Illinois	23276	53.93538409	42.79273032
May4	Illinois	28879	65.35198587	63.31613559
OSU611	Illinois	24002	26.62694775	13.16984516
OSU622	Illinois	97198	23.2576802	30.50002989
OSU624	Illinois	65298	38.88480505	89.69241346
OSU626	Illinois	20919	37.39662508	35.4464638
OSU628	Illinois	56784	20.31558185	29.37203324
OSU634	Illinois	1427900	13.33923944	45.14933939
OSU638	Illinois	76797	21.19874474	61.43151791
P59	Illinois	134568	17.28642768	26.48329049
P35	Illinois	27513	23.85054338	3.871644646
PRD1	Illinois	116751	71.87261779	129.2258026
PRD10	Illinois	699568	75.15180797	14.05033778
PRD9	Illinois	90571	73.43630964	127.813535

Table 3.3 Sequencing statistics for ancient dog samples included in nuclear DNA analyses. 'Sites' corresponds to the number of sites out of the ~6.21 million sites that were called in each sample.

Sample ID	Extraction lab	Number of reads	Endogenous content	Nuclear coverage	Sites
AL2135	Oxford/Illinois	7,366,980	3.37	0.006	16,869
AL2748	Oxford	18,128,973	6.694	0.023	98,361
AL3198	Oxford	18,711,277	2.816	0.001	38,806
AL2772	Oxford	13,783,057	4.596	0.013	59,871
AL3226	Oxford	36,476,977	4.5	0.031	119,689
AL3223	Oxford	125,598,030	16.891	0.452	1,806,093
AL3194	Oxford	265,763,856	38.336	1.921	4,661,948

3.5.2 Data processing

Ancient DNA data processing was performed by Laurent Frantz. Raw reads were filtered allowing one mismatch to the indices used in library preparation. Adapter sequences were removed using AdapterRemoval (Lindgreen, 2012). Reads were aligned using Burrows-Wheeler Aligner (BWA, v0.7.5ar405) (Li and Durbin, 2009) to CanFam3.1 (Schubert et al., 2012). PCR duplicates were removed and BAM files from different sequencing lanes were merged using the MergeSamFiles tool from Picard v1.129 (http://broadinstitute.github.io/picard/).

To accommodate for the low coverage of the ancient American dog nuclear genome sequences, genotypes were called by randomly sampling a single read of 20 bp minimum and with a mapping quality (MAQ) and base quality (BQ) of at least 30 at each covered position in the genome, excluding bases within 5 bp of the start and end of a read (Green et al., 2010; Haak et al., 2015; Skoglund et al., 2016).

The Newgrange dog (Frantz et al., 2016; Section 3.6) was genotyped in the same way as modern data (except for 5 bp at start and end of a read; see below).

For the mtDNA, we generated majority consensus sequences for all samples that had at least 3× average coverage (71 samples; Table 3.2) excluding bases within 5 bp of the start and end of a read. Molecular damage was assessed using MapDamage2.0 using default parameters (Jónsson et al., 2013). Most samples displayed clear signs of deamination. Samples AL3231 and AL2696 displayed limited deamination pat-

terns consistent with these being from a relatively recent time period (AD 1000-1400; Table 3.1).

3.6 Publicly available data

Raw reads/bam files for 47 canid genomes (Fan et al., 2016; Frantz et al., 2016; Freedman et al., 2014; Wang et al., 2016) were downloaded from NCBI or DoGSD (Bai et al., 2015). Samples downloaded from NCBI were aligned to the CanFam3.1 reference genome using BWA mem (Li and Durbin, 2009). We computed depth of coverage for each sample using bedtools (Quinlan and Hall, 2010). These genomes were chosen due to their high coverage and geographic spread covering North and South American, East Asian, and Western European (African, Indian and European) dogs, as well as Eurasian and American grey wolves and coyotes and an outgroup (*Lycalopex culpaeus*; Andean fox).

Table 3.4 Summary information (coverage, accession, etc.) of modern whole genomes analysed in this chapter

Accession	Coverage	Sample Id	Population code	Meta population	Info	Country
PRJEB7788	1.3	W_Taimyr	Taimyr	Ancient Wolf	Wolf (ancient)	Taimyr Peninsula (Siberia)
ERS403439	124.2	C_24T	CTVT	CTVT	CTVT	Australia
ERS403440	24.5	D_24H	DCTVT	CTVT Host	Aboriginal camp dog	Australia
ERS1870073	96.2	C_608T	CTVT	CTVT	CTVT	The Gambia
ERS1870074	47.7	D_608H	DCTVT	CTVT Host	Local Gambian dog	The Gambia
ERS1870075	94.3	C_609T	CTVT	CTVT	CTVT	The Gambia
ERS1870076	43.7	D_609H	DCTVT	CTVT Host	Local Gambian dog	The Gambia
ERS403441	95.1	C_79T	CTVT	CTVT	CTVT	Brazil
ERS403442	23.7	D_79H	DCTVT	CTVT Host	American cocker spaniel	Brazil
ERS747137	13.4	D_AHusky91	DAL	Arctic Dogs	Alaskan Husky	Alaska, USA
SRS1129580	15.8	D_Mal68	DMA	Arctic Dogs	Alaskan Malamute	Alaska, USA
DoGSD	6.3	D_Basenji	BAS	African Dogs	Basenji	Africa
SRS661477	32.8	C_Cal	COY	Coyotes	Coyote	California, USA
DoGSD	6.7	D_Dingo	DIN	Dingo	Dingo	Australia
DoGSD	14.9	D_Tibet3	DTI	East Asian Dogs	Village Dog	Tibet
DoGSD	15.5	D_Tibet4	DTI	East Asian Dogs	Village Dog	Tibet
SRS1129835	15.5	D_Green	DGL	Arctic Dogs	Greenland Sledge Dog	Unknown
DoGSD	15.9	D_GerShep3	DGS	European Dogs	Germany Shepherd	Germany
DoGSD	16.3	D_GerShep6	DGS	European Dogs	Germany Shepherd	Germany
SRS984799	61.8	D_Husky	DHU	Arctic Dogs	Husky	Unknown
SRS520064	10.0	D_India168	DID	Asian Dogs	Village Dog	India
SRS520065	15.5	D_India60	DID	Asian Dogs	Village Dog	India
SRS661487	43.6	W_India	WME	Eurasian Wolf	Wolf	India
SRS661488	27.3	W_Iran	WME	Eurasian Wolf	Wolf	Iran
SRS520071	9.9	D_Leb79	DLB	European Dogs	Village Dog	Lebanon
SRS520072	10.5	D_Leb85	DLB	European Dogs	Village Dog Lebanon	Lebanon
SRS523207	12.1	AndeanFox	OUT	Outgroup	Andean Fox	Andes
DoGSD	16.3	D_China8	DCH	East Asian Dogs	Village Dog	China
DoGSD	16.8	D_China9	DCH	East Asian Dogs	Village Dog	China
DoGSD	9.2	W_Mongo	WAS	Eurasian Wolf	Wolf	Mongolia
DoGSD	11.3	W_Altai	WAS	Eurasian Wolf	Wolf	Altai
SRS1135618	16.3	D_Mex	DME	American Dogs	Mexican Hairless Dog	Mexico
SRS661479	8.4	C_MidW	COY	Coyotes	Coyote	Midwest, USA
SRS661490	24.7	W_Mex1	WAM	American Wolf	Wolf Mexico	Mexico
SRS520075	15.2	D_Na8	DNA	African Dogs	Village Dog	Namibia
SRS520075	7.6	D_Na89	DNA	African Dogs	Village Dog	Namibia
ERS1097373	33.6	D_NGDG	DAE	European Dogs	Newgrange Dog (Ancient)	Ireland
SRS1135617	17.5	D_Peru	DPU	American Dogs	Peruvian Naked Dog	Peru
SRS520079	18.6	D_Port61	DEU	European Dogs	Village Dog	Portugal
SRS520080	16.2	D_Portt71	DEU	European Dogs	Village Dog	Portugal
SRS661492	31.4	W_Port	WEU	Eurasian Wolf	Wolf	Portugal
SRS520081	12.9	D_Qatar27	DQA	Asian Dogs	Village Dog	Qatar
SRS520082	7.5	D_Qatar5	DQA	Asian Dogs	Village Dog	Qatar
SRS1135615	13.9	D_Husky89	DHU	Arctic Dogs	Siberian Husky	Siberia
SRS1129824	16.6	D_SLaika	DSL	Spitz Dogs	Siberian Laika	Siberia
SRS661495	23.5	W_Spa	WEU	Eurasian Wolf	Wolf	Spain
DoGSD	13.7	D_TMastif4	DTM	East Asian Dogs	Tibetan Mastiff	Tibet
DoGSD	15.1	D_TMastif5	DTM	East Asian Dogs	Tibetan Mastiff	Tibet
SRS520084	10.9	D_Viet21	DVN	East Asian Dogs	Village Dog	Vietnam
SRS520088	10.6	D_Viet59	DVN	East Asian Dogs	Village Dog	Vietnam
SRS661496	26.8	W_Yellow1	WAM	American Wolf	Wolf	Yellowstone, USA
SRS661497	25.7	W_Yellow2	WAM	American Wolf	Wolf	Yellowstone, USA

In addition, we obtained data from two publicly available CTVT genomes (Murchison et al., 2014), as well as genomic data from the ancient Taimyr wolf (Skoglund et al., 2015; Section 3.3.2) and a Late Neolithic Irish dog from Newgrange (Frantz et al., 2016; Section 1.3). Whole genomes from two dog samples in Germany dating to the Early and End Neolithic published by Botigué et al. (2017) have not been included in the ensuing analyses as these were not available at the time.

We additionally obtained data from 5,406 modern dogs that were genotyped on the semi-custom CanineHD SNP array (~185K SNPs) developed by Shannon et al. (2015).

3.7 Genotyping

3.7.1 Normal dog samples

We used SAMtools 'mpileup' v0.1.19 (Li et al., 2009) to call genotypes with default settings. Pileup files were filtered, for all samples, using the following criteria:

- Minimum depth of coverage (DoC) ≥ 6
- Exclude all sites in region of high DoC (top 5%)
- Exclude all sites within 3 bp of an indel
- Only bases with quality $\geq$30 within reads with mapping quality ≥ 30 were used.
- Minimum fraction of reads supporting heterozygous (variant allele frequency [VAF] >= 0.3) - all sites that did not pass this criteria ($0 < \text{VAF} < 0.3$) were coded as missing (N).
- In the high coverage ancient sample (Newgrange dog) we also discarded the first and last 5 bp of each read for genotype calling, to avoid incorporating errors from deaminated sites (see above).

The Taimyr wolf (Section 3.6; Skoglund et al., 2015) was processed using the same random read approach used for other ancient samples (see Section 3.5.2).

3.7.2 CTVT samples

Ancestry analyses were performed using data from two previously described CTVT genomes, 24T and 79T, (Murchison et al., 2014) in order to determine the phylogenetic placement of the CTVT founder dog within a cohort of modern and ancient dogs. CTVT genomes carry two types of genetic variation: germline variation inherited by the CTVT founder dog, and somatic variation acquired during the somatic evolution of the CTVT clone. The goal of this specific analysis was to identify CTVT germline variation, and to use this to represent the CTVT founder individual in a phylogenetic analysis.

We generated a list of callable sites in CTVT using the criteria outlined in Section 3.7.1. Because CTVT is a cancer, and to limit the impact of somatic mutations, we confined our genotyping analysis to SNPs mapping to genomic regions in CTVT that have retained both parental chromosomal copies (i.e. germline diploidy), as previously described (Murchison et al., 2014). Sites were further filtered to retain only those sites in which the variant allele fraction (VAF) for a non-reference allele was ≥ 0.1 in at least one CTVT tumour, and for which no more than two nucleotides were detected at VAF ≥ 0.1 (i.e. multi-allelic sites were rejected). These sites were defined as single nucleotide variant (SNV) candidates.

CTVT tumour tissue biopsies contain both CTVT and host cells. The latter derive from stromal, immune and blood vessel components. Thus DNA derived from CTVT tumours is an amalgam of CTVT and matched host DNA. In order to identify and exclude SNVs derived exclusively from the matched host, as well as to correctly genotype alleles shared between CTVT and matched hosts, we took the following approach. Sites identified as SNV candidates were genotyped in genomes 24H and 79H, the matched hosts for tumours 24T and 79T (Murchison et al., 2014). The genotypes of 24H and 79H were inferred using the following VAF thresholds: homozygous reference if VAF ≤ 0.2, heterozygous if VAF = [0.2-0.8], homozygous alternative if VAF ≥ 0.8.

Using the known host contamination fractions for 24T and 79T (Murchison et al., 2014), we used the following VAF thresholds to genotype SNV candidates in CTVT cells. SNVs that were homozygous reference in the matched host were genotyped in CTVT using the following VAF thresholds: homozygous reference for VAF$<$0.2,

heterozygous if VAF = [0.2-0.6], homozygous alternative if VAF>0.6. SNVs that were heterozygous in the matched host were genotyped in CTVT using the following VAF thresholds: homozygous reference if VAF<0.3, heterozygous if VAF = [0.3-0.7], homozygous alternative if VAF>0.7. SNVs that were homozygous alternative in the matched host were genotyped in CTVT using the following VAF thresholds: homozygous reference if VAF<0.4, heterozygous if VAF =[0.4-0.8], homozygous alternative if VAF>0.8. CTVT SNVs were further processed as described below in Section 3.7.3.

3.7.3 Ascertainment panel

Genotypes from all genome-wide samples were then merged using bedtools (Quinlan and Hall, 2010). Ascertainment was done without outgroups (but including coyote). We selected all bi-allelic markers excluding sites that (*i*) were heterozygous in only one sample (required a minimum of two chromosomes in our set of samples to carry the derived allele); in the case of sites only variable in CTVT we required the two CTVT genomes (24T and 79T) to be homozygous to limit the inclusion of somatic mutations into the list of SNPs, (*ii*) sites that were not covered in our outgroup (Andean fox), and (*iii*) sites with more than 20% missing data across samples. All low coverage ancient samples were excluded from this step (Table 3.3). This resulted in ~6.21 million high quality SNPs. We excluded all sites that were outside of the germline diploid region in CTVT (Murchison et al., 2014). Singleton SNPs called exclusively in CTVT genomes and not found in any other canid genome were excluded. This resulted in ~2.03 million SNPs, including ~600K transversions.

3.8 Mitochondrial DNA analyses

3.8.1 RAxML

For phylogenetic analysis, we used all samples with at least 3× average mitochondrial coverage and consensus sequences with at least 80% coverage across the entire mitochondrial genome were considered for further analysis (Table 3.2). We also included previously published ancient and modern mtDNA genomes (Thalmann et al., 2013). This data set contains representative samples of all four major dog mitochondrial haplogroups (A, B, C, D) including 3 ancient American dogs. We aligned the data using mafft v7.2 (Katoh and Standley, 2013). We built a maximum

likelihood (ML) tree in RAxML (Stamatakis, 2006), with 100 bootstrap replicates, using a General Time Reversible (GTR) substitution model with a gamma (G) model of site heterogeneity.

We assessed whether previous studies relying on the mtDNA control region (see Section 2.3.1) were able to identify the pre-contact monophyletic clade we have identified here.To do so, we extracted the control regions from all samples that overlapped with the fragments analysed in (Leonard et al., 2002) and (van Asch et al., 2013) (605 bp in total), filtering out samples with more than 10% missing data and a maximum likelihood tree with RAxML.

Lastly, we expanded our MT genome sample size to assess whether the mtDNA haplogroup that we had identified in pre-contact dogs exists in modern America dogs. To do so we included 942 additional mitogenomes from a global sample of dogs, including CTVT host genomes as well as 169 village and breed dogs that were sampled in North and South America (Björnerfeldt et al., 2006; Pang et al., 2009; Webb and Allard, 2009; Zhao and Liu, 2016). Accession number of all additional samples can be found in Appendix 1: Supplementary file 3.3. We combined this data with the mtDNA genomes analysed above and built a ML tree, with 100 bootstrap replicates using a GTR+G model as implemented in RAxML (Stamatakis, 2006).

3.8.2 Bayesian phylogenetic analysis

We used BEAST v1.8.4 (Drummond et al., 2012) to calibrate the evolutionary rate of our canid data set. We restricted this analysis to sequences with at least 10$\times$ average mitochondrial genome coverage (Table 3.2). The mitogenome was partitioned into four categories (tRNA, rRNA, control region and coding sequence). We fitted a separate substitution model to each partition: tRNA (HKY+I), rRNA (TN93+G), control region (HKY+G+I) and coding sequence (SDR06) as selected by Akaike information criteria (AIC) using PartitionFinder (Lanfear et al., 2017). The same tree was used for all four partitions. The age of archaeological samples was used as prior (uniform distribution of tip age). We used a Bayesian Skyline prior (Drummond et al., 2005) (group size parameter = 10) and a strict molecular clock as in Thalmann et al. (2013). An uncorrelated clock was also tested and did not result in noticeable changes. We ran 50 million Markov chain Monte Carlo (MCMC) chains and sampled tree

parameters every 5,000 iterations. Convergence was evaluated with Tracer v1.6.0 (ESS for each parameter≥100). Trees were summarised using Maximum Clade Credibility as implemented in TreeAnnotator v1.8.4 (10% burn-in).

3.9 Nuclear analyses

3.9.1 Nomenclature

Hereafter, I will refer to the ancient dog individuals sampled in North America as pre-contact dogs (PCD) although I do not formally define this group until Sections 3.11.1 and 3.11.2 based the results of phylogenetic analyses.

3.9.2 PCA

Principal Component Analysis (PCA) was performed using smartpca (Patterson et al., 2006) on a set of 2.03 million SNPs using the following projections and data sets:

(i) Pre-contact samples projected

(ii) Dog samples only with pre-contact samples projected

(iii) Dog samples only with pre-contact and CTVT samples projected

For PCD we used all 7 samples for which we could call at least 10,000 sites (minimum number of sites suggested for ancient DNA analysis (Allentoft et al., 2015)). We used all available sites (sites covered in at least 1 ancient sample; ~1.5 million SNPs) to compute the eigenvectors and then projected PCD onto that space. We also projected CTVT to ensure that their placement was not an artefact of somatic mutations.

3.9.3 Neighbour-joining tree

An Identity By State (IBS) matrix was computed using plink v1.9 (Purcell et al., 2007) and using all 2.03 million SNPs. This matrix was used to build a neighbour-joining tree (NJ) using the R package "ape" (Paradis et al., 2004). We built a phylogeny using nuclear genotypes with MrBayes 3.2 (Ronquist and Huelsenbeck, 2003). To do so we used PGDSpider 2.0.9.2 (Lischer and Excoffier, 2012) to build a Nexus file with discrete SNP format (0=reference, 1 = heterozygous, 2 = homozygous alternative).

We used the Mkv model (Stamatakis, 2006) implemented in MrBayes (ordered character), which provides a likelihood framework for data sets that contain only variable characters. We also imposed a minimum distance of 10Kb between SNPs to limit the influence of linkage disequilibrium (LD) and lastly included only PCD samples with higher coverage (AL3194 and AL3223; Table 3.3; ~30K SNP total).

We ran two independent runs of four MCMC chains with two million samples. Trees were summarised discarding 25% as burn-in. To limit biases from missing data we limited this analysis to transversions that were covered in 90% of our samples. Convergence was assessed by ensuring that average standard deviation of split frequencies was below 0.01 and that the potential scale reduction factor was close to 1 for all parameters.

3.9.4 f_3 statistics

We computed outgroup f_3-statistics (3-population tests) (Patterson et al., 2012) using ADMIXTOOLS to explore the relationship between pre-contact dogs and present-day dog populations. We computed f_3(Pre-contact dogs, X; outgroup) where X is any other dog population (Table 3.4), to quantify the amount of genetic drift shared between pre-contact dogs and other dogs, using only transversions.For this analysis, we used only the higher coverage PCD samples (AL3194 and AL3223; Tables 3.1 and 3.3), with ~1.9× and ~0.5× coverage, respectively.

3.9.5 TreeMix

To relate the population history of pre-contact dogs and CTVT to other dog populations and to test the topology suggested by phylogenetics and f_3 statistics analyses we used TreeMix (Pickrell and Pritchard, 2012) to reconstruct admixture graphs. We only used 3 representatives from each major dog group:

- West Eurasian dogs: Portuguese village dogs (DEU), German Shepherd (DGS)
- East Asian dogs: Vietnamese village dogs (DVN) because they lack admixture from European dogs (see above) and Tibetan village dogs (DTI)
- Pre-contact dogs (PCD): Port au Choix (AL3194) and Weyanoke Old Town (AL3223) samples (Table 3.1)
- Arctic dogs: Malamute (DMA) and Greenland dogs (DGL)

- CTVT: 24T and 79T
- Eurasian wolves (WEU) from Spain and Portugal
- North American wolves (WAM) from Yellowstone
- Coyotes (COY) as an outgroup

Malamute and Greenland dogs were chosen as they are the least admixed with Western dogs. We only used transversions in order to limit the effect of DNA damage on the analysis and only used sites that were covered in all samples (~60,000 SNPs).

3.9.6 qpGraph

QpGraph (Patterson et al., 2012) was used to fit admixture graphs to nine populations (see below) representing PCD, CTVT, and each of the three major dog groups, plus wolves and coyotes.

- West Eurasian dogs: Portuguese village dogs (DEU)
- East Asian dogs: Vietnamese village dogs (DVN)
- Pre-contact dogs (PCD): AL3194 (Port au Choix) and AL3223 (Weyanoke Old Town; Table3.1)
- CTVT: 24T and 79T
- Arctic dogs: Alaskan malamute (DMA)
- Eurasian wolves (WEU) from Spain and Portugal
- North American wolves (WAM) from Yellowstone
- Coyotes (COY) from California
- Andean fox (OUT) as the outgroup

We only used transversions in order to limit the effect of DNA damage on the analysis. This resulted in 600,991 high quality SNPs.

To explore the space of all possible admixture graphs we implemented a heuristic search algorithm. Given an outgroup with which to root the graph, a stepwise addition order algorithm was used for adding leaf nodes to the graph. At each step, insertion of a new node was tested at all branches of the graph, except the outgroup branch. Where a node could not be inserted without producing f_4 outliers

(i.e. $|Z| \geq 3$) then all possible admixture combinations were also attempted. If a node could not be inserted via either approach, that sub-graph was discarded. If the node was successfully inserted, the remaining nodes were recursively inserted into that graph. All possible starting node orders were attempted to ensure full coverage of the graph space.

As the number of possible graphs grows super-exponentially with each additional leaf node, we initially excluded CTVT from the search space and looked for models which fit the remaining eight populations. We fitted 480,166 unique admixture graphs for these populations and recorded the 892 graphs that left no f_4 outliers (i.e. $|Z| < 3$). We then fitted a further 309,525 unique models, testing all possible insertions of CTVT into the 892 eight-population graphs, and recorded the 1,655 graphs that left no f_4 outliers.

TreeMix analysis was also performed using the same nine populations, with six admixture edges (the maximum number seen in the qpGraph analyses; see Section 3.9.6). The most plausible qpGraph model (Fig. 3.14) was chosen by comparing all fitted models to the neighbour-joining tree, Bayesian tree, *D*-statistics analyses and a TreeMix tree with the same sampling.

3.9.7 *D*-statistics

D-statistics were computed using ADMIXTOOLS (Patterson et al., 2012). For these analyses we used only the two higher coverage PCD genomes (AL3194 and AL3223; Table 3.3), with average genome coverage ~1.9× and ~0.5× respectively, except where explicitly stated (e.g. Koster dog AL2135; see below). Results are based on all 2.03 million SNPs. The Andean fox was used as an outgroup for these analyses. We obtained standard errors using a block jackknife procedure over 1 megabase (1Mb) blocks in the genome.

- For each pair of samples (Pop 3, Pop 4) we computed *D*(Outgroup, PCD, Pop 3, Pop 4) where Population 3 (Pop 3) was fixed as either European dogs, Asian dogs, Arctic dogs, or CTVT and Population 4 (Pop 4) represented any other positive sample. Positive values imply that pre-contact dogs shared more derived alleles with the population on the y axis, while negative values imply that PCD are closer to other dog populations.

- To test for admixture from wild North American canids into PCD (AL3194, AL3223), we computed *D*(Outgroup, Coyote or North American Wolf, Pop 3, Pop 4) where Pop 3 and Pop 4 could represent any possible pair of populations.
- To test for additional admixture from wild canids into PCD we ran *D*(Outgroup, Coyote, AL3194, AL3223) and *D*(Outgroup, American Wolf, AL3194, AL3223).
- We computed *D*(Outgroup, Coyote, CTVT, Pop 4) and *D*(Outgroup, American Wolf, CTVT, Pop 4) where Pop 4 could represent any other dog population to determine whether there was detectable admixture from coyote and American wolf populations into the CTVT founder dog specifically.
- We used whole genome data to assess admixture from the ancient Taimyr wolf into PCD, Arctic dogs and CTVT by computing *D*-statistics of the form *D*(Outgroup, Taimyr, PCD or Arctic dogs, European, Asian, or Arctic dogs).
- We tested for admixture from European dogs into East Asian dogs. Following Shannon et al. (2015) and Frantz et al. (2016), we used Vietnamese village dogs as the reference East Asian population to test for by computing *D*(Outgroup, Portugal, Vietnam, Pop 4) (Table 3.9).
- To test for coyote admixture in the Koster dog (AL2135), we computed *D*(Outgroup, North American wolf or coyote or Taimyr wolf, AL2135, AL3194). This analysis was restricted to AL3194 as it was the highest coverage PCD sample available (Table 3.3).
- *D*-statistics were used to test for admixture from PCD into Arctic breeds since their MRCA by computing *D*(Outgroup, PCD (AL3194), Arctic dogs, Arctic dogs) (Table 3.10).
- Lastly, to determine whether admixture from Eurasian dogs could affect the result of the previous *D*-statistic test, we quantified *D*(Outgroup, Asian or European Dogs, Arctic dogs, Arctic dogs).

3.9.8 f_4 ratio estimation

We used SNP array data obtained from Shannon et al. (2015) to assess the degree to which modern dog populations in the Americas have retained ancestry from pre-contact dogs. This SNP panel contained 28 genotyped populations from North and South America, such as Peruvian village dogs, Alaskan village dogs and Carolina dogs (Table 3.5 for the full list).

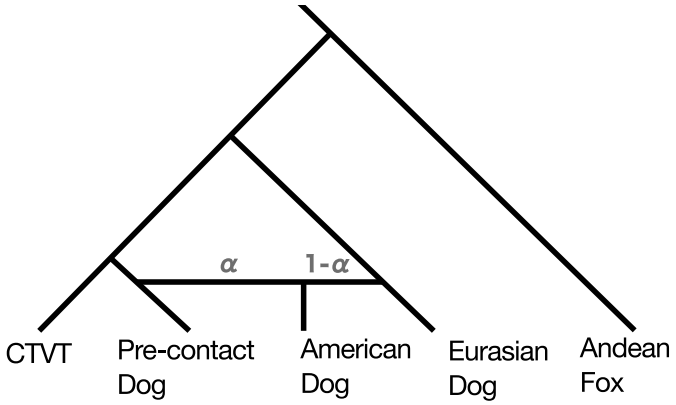


Fig. 3.2. Schematic representation of the assumed phylogeny for the f_4 ratio test used to estimate pre-contact ancestry into modern North American dogs. Here, alpha represents the degree of ancestry from pre-contact dogs. Figure adapted from Ní Leathlobhair et al. (2018).

f_4 ratios were computed using ADMIXTOOLS (Patterson et al., 2012; Reich et al., 2009) to estimate admixture proportions (α) from pre-contact dogs into these populations by computing:

$$\alpha = f_4(A, O; X, C) \div f_4(A, O; B, C)$$

Where A is CTVT, O is the Andean fox (outgroup), B is PCD (AL3194 and AL3223), C is any European or East Asian population (Table 3.5) and X is any American dog (Table 3.5 and Fig. 3.2). We computed α for all combinations of European/Asian and modern American dog populations. Jackknifing was performed with a block size of 1 centiMorgan (cM). To test whether Alaskan village dogs interbred with pre-contact dogs we computed every possible combination of the same f_4 ratio as before but using only the Arctic subset of American dogs. Outgroup f_3 statistics were also used to assess the degree of shared drift between various populations available on the SNP array and PCD.

3.9.9 ADMIXTURE

To further assess this we used ADMIXTURE (Alexander et al., 2009) on a subset of the SNP array samples including all modern North American populations as well as Arctic dogs, "basal" breeds (Larson et al., 2012) and selected European and Asian

Table 3.5 Sample codes used as population names for SNP array analysis

Sample code	Dog type	Type group
AED	American Eskimo Dog	American Dogs
AM	Alaskan Malamute	Northern Dogs
APBT	American Pit Bull Terrier	American Dogs
AST	American Staffordshire Terrier	American Dogs
BAS	Basenji	African Dogs
BAS2	Basenji	African Dogs
BEA	Beagle	European Dogs
BOX	Boxer	European Dogs
CBR	Chesapeake Bay Retriever	American Dogs
CC	Chow Chow	East Asian Dogs
CD	Carolina Dog	American Dogs
CHI	Chihuahua	American Dogs
CLD	Catahoula Leopard Dog	American Dogs
COO	Chinook	American Dogs
COY	Coyote	Coyotes
CSP	Chinese Shar-pei	East Asian Dogs
CTVT	CTVT	CTVT
DAE	Ancient European	European Dogs
DAL	Alaskan Husky	Northern Dogs
DCH	Chinese Village Dog	East Asian Dogs
DEU	European Village Dog	European Dogs
DGL	Greenland Dogs	Northern Dogs
DGS	German Shepard	European Dogs
DHU	Husky	Northern Dogs
DID	Indian Village Dog	Asian Dogs
DIN	Dingo	Dingo
DLB	Lebanese Village Dog	European Dogs
DMA	Malamute	Northern Dogs
DME	Mexican Hairless Dog	American Dogs
DNA	Namibian Village Dog	African Dogs
DPU	Peruvian Hairless Dog	American Dogs
DQA	Qatari Village Dogs	Asian Dogs
DSL	Siberian Laika	Northern Dogs
DTI	Tibetan Village Dog	East Asian Dogs
DTM	Tibetan Mastiff	East Asian Dogs
DVN	Vietnamese Village Dog	East Asian Dogs
EUR	Eurasier	Northern Dogs
FS	Finnish Spitz	Northern Dogs
GSD	Greenland Sledge Dog	Northern Dogs
NEW	Newfoundland	American Dogs
NGSD	New Guinea Singing Dog	East Asian Dogs
NSDTR	Nova Scotia Duck Tolling Retriever	American Dogs
PCD	Pre-Colombian Dogs	Pre-Colombian Dogs
PIO	Peruvian Inca Orchid	American Dogs
SAM	Samoyed	Northern Dogs
SH	Siberian Husky	Northern Dogs
VDB	Village Dog Belize	American Dogs
VDB2	Village Dog Brazil	American Dogs
VDC	Village Dog Colombia	American Dogs
VDCR	Village Dog Costa Rica	American Dogs
VDDR	Village Dog Dominican Republic	American Dogs
VDH	Village Dog Honduras	American Dogs
VDP	Village Dog Panama	American Dogs
VDPA	Village Dog Peru-Arequipa	American Dogs
VDPC	Village Dog Peru-Cusco	American Dogs
VDPI	Village Dog Peru-Ica	American Dogs
VDPL	Village Dog Peru-Loreto	American Dogs
VDPP	Village Dog Peru-Puno	American Dogs
VDPR	Village Dog Puerto Rico	American Dogs
VDUA	Village Dog US-Alaska	American Dogs
VDIC	Village Dog India-Chennai	Asian Dogs
VDID	Village Dog India-Dehli	Asian Dogs
VDIH	Village Dog India-Hazaribagh	Asian Dogs
VDIM	Village Dog India-Mumbai	Asian Dogs
VDIO	Village Dog India-Orissa	Asian Dogs
VDIB	Village Dog Indonesia-Borneo	East Asian Dogs
VDIJ	Village Dog Indonesia-Jakarta	East Asian Dogs
VDPNGEH	Village Dog Papua New Guinea-East Highlands	East Asian Dogs
VDPNGPM	Village Dog Papua New Guinea-Port Moresby	East Asian Dogs
VDVCB	Village Dog Vietnam-Cao Bang	East Asian Dogs
VDVHG	Village Dog Vietnam-Ha Giang	East Asian Dogs
VDVLS	Village Dog Vietnam-Lang Son	East Asian Dogs
VDVLC	Village Dog Vietnam-Lao Cai	East Asian Dogs
WAM	American Wolf	American Wolf
WAS	Asian Wolf	Eurasian Wolf
WEU	European Wolf	Eurasian Wolf
WME	Middle-east Wolf	Eurasian Wolf
XOL	Xoloitzcuintli	American Dogs
TAI	Taimyr	Ancient Wolf
OUT	Andean Fox	Outgroup

populations (e.g. Boxer and Chow-Chow). K = 4 was selected as the best K value based on 10 fold cross validation. We tried to separate PCD/Arctic ancestry with higher K values.

3.10 CTVT mutation rate analysis

3.10.1 Experimental design

Our goal was to estimate the CTVT somatic mutation rate and to use this to estimate the time at which CTVT originated. To do this, we collected biopsies from a pair of CTVT tumours involved in a naturally occurring direct transmission event, and identified mutations that had arisen during the known transmission time interval to define a somatic mutation rate. We then estimated the number of somatic mutations in the entire CTVT lineage and applied our somatic mutation rate to estimate the time of CTVT origin. Previous estimates of CTVT time of origin are described in Chapter 1 (Section 1.2.1.7).

3.10.2 Sample collection

This project was approved by the Department of Veterinary Medicine, University of Cambridge, Ethics and Welfare Committee (reference CR174). A 1-2 mm^3 biopsy was sampled from Dog 609's vaginal tumour (609T). A 1-2 mm^3 biopsy was sampled from one of Dog 608's skin tumours on the same day (608T). Biopsies were also collected from 609H and 608H host tissues (ovary and testis, respectively). CTVT diagnosis was confirmed as previously described (Strakova et al., 2016).

3.10.3 Sample case histories

Dog 609H was a mixed-breed free-ranging dog from the Gambia with an approximately 31 cm^3 vaginal CTVT tumour. Her ten-month-old male pup, Dog 608, had several CTVT tumours on the ventral skin. This unusual CTVT presentation in Dog 608 suggested that CTVT cells were likely engrafted from its mother's vaginal tumour during parturition.

Fig. 3.3. Samples sequenced in this study.
Sample 608T is a CTVT biopsy from the skin of a 10-month old puppy. This tumour was likely transmitted during birth and engrafted from the mother's vaginal tumour (609T). Tumour (608T, 609T) and host (608H, 609H) DNA was sequenced from the two individuals shown. Images are reproduced with permission from Michael Meyer.

3.10.4 Variant calling

3.10.4.1 Extraction and filtering

We used Somatypus (https://github.com/baezortega/somatypus), a Platypus (Rimmer et al., 2014) based variant calling and genotyping pipeline, to identify SNVs and indels. In order to make an initial call, SNVs were required to have $\geq$3 supporting reads in at least one of the four sequenced CTVT or matched host samples (608T, 608H, 609T, 609H). Indels were input to GATK Realigner Target Creator (McKenna et al., 2010) for local realignment and SNVs were re-called from realigned genomes.

The following in-built Platypus flags were used to exclude SNVs at two stages, before and after genotyping: badReads, MQ, QD, strandBias, SC.
The following post-processing filters were also implemented:

- Strand bias filter. For each SNV, the total coverage, as well as forward and reverse strand read support were extracted. For low total coverage positions ($\leq$10 reads across all four samples), we discarded calls with less than two supporting reads in either forward or reverse direction. For high total coverage positions ($>$10 reads across all four samples), we discarded calls with less than 20% support on either the forward or reverse sequencing strands.

- Simple repeat filter. SNVs within simple repeats, as defined by the UCSC table browser (CanFam3.1), were excluded.

- Extreme depth filter. SNVs within regions of high read depth were also excluded. To detect high read depth (HRD) regions we first generated BigWig coverage files from matched normal whole genome sequence data files (608H, 609H). We then identified areas with coverage 12 standard deviations higher than the mean read coverage, on a chromosome by chromosome basis. Common intervals between normal samples were identified using bedtools multiinter (Quinlan and Hall, 2010) and were merged using bedtools merge. The maximum allowed distance between regions to be merged was 250 bp. HRD regions spanning less than 500 bp were excluded. Any HRD region that overlapped with gene regions as defined by the UCSC table browser (CanFam3.1, Genes and Gene Predictions, Ensembl Genes) was excluded.

- Low VAF filter. SNVs with VAF>0 and VAF<0.2 in both 608H and 609H were discarded if (*i*) they were not detected in 608T or 609T or (*ii*) they were found with VAF>0 and VAF<0.1 in either or both of 608T and 609T.
- Regions filter. SNVs occurring in the mitochondrial genome or on unassigned scaffolds were excluded. In addition, to avoid problems caused by variable coverage in hosts, SNV analysis was restricted to autosomes.

3.10.4.2 Germline and consensus filtering

SNVs identified in 608T and 609T will belong to one or more of the following categories:

(i) contaminating germline SNPs from matched host,

(ii) germline SNPs inherited by the CTVT founder dog, and

(iii) somatic mutation SNVs.

In order to enrich for somatic mutations, we filtered our candidate SNVs against a panel of 28,812,954 canid germline SNPs. We excluded any genomic site that was reported in any of the following variant catalogues:

- 608H and 609H (sites with $\geq$5 reads coverage and $\geq$2 reads supporting a non reference allele were considered SNVs)
- The Variant and Systematic Error Catalogue (VSEC) (Decker et al., 2015)
- The CanineHD 170K SNP array (Vaysse et al., 2011)
- The ascertainment panel generated in this study prior to incorporating CTVT samples (see Section 3.7.3)
- A complete genome from a Greenland sledge dog (Wang et al., 2016; Table 3.4) included in the ascertainment panel was additionally genotyped. This provided additional SNVs beyond those in the ascertainment panel, as the ascertainment panel excluded SNVs that were found on only one chromosome; thus SNVs that were exclusively found in the Greenland sledge dog individual and CTVT would not have been included in the ascertainment panel that we filtered against, but were excluded in this step (see Section 3.7.1).

Next, we further filtered the remaining SNVs using the following criteria:

- We retained only those SNVs that had that had $\geq$2 reads supporting the variant all with minimum base quality of 20 and minimum mapping quality of 35, in at least one of the two tumours using the alleleCount tool https://github.com/cancerit/alleleCount.
- We retained only those SNVs that were identified by GATK Haplotype Caller. The GATK engine was restricted processing candidate loci.
- We required that the matched host must have coverage of at least 20 reads total at the candidate SNV position. Candidate SNVs that did not reach this threshold in one or both tumours were discarded.
- We discarded SNVs with $\geq$10 reads total (regardless of whether they supported the variant) with base quality $<$20 and mapping quality $<$35 in matched hosts

1,934,103 and 1,934,125 tumour-only SNVs remained in 608T and 609T respectively after these steps; 1,933,897 of these were shared by 608T and 609T. Of the SNVs in this set that mapped to genomic regions retaining both parental copies, almost all SNVs were heterozygous. Thus, the majority of these SNVs are likely to be somatic; however, some germline variation that was present in the CTVT founder dog, but that is not represented in the germline panel used here likely still remains. It is also likely that some somatic mutations, which occurred in the same sites as germline SNVs represented in our panel, have been removed.

3.10.4.3 Tumour-unique SNVs

We next filtered tumour-only SNVs, as defined above, for those unique to either 608T or 609T. In order to be considered unique to a single tumour, a variant was required to be present with $\geq$2 supporting reads in only one tumour, with minimum base quality of 20 and minimum mapping quality of 35 for those reads supporting the variant. This method yielded 206 tumour-unique SNVs in 608T and 228 tumour-unique SNVs in 609T.

3.10.5 Mutational spectrum

Each tumour-only SNV was classified as one of six possible mutation types in the pyrimidine context (C>A, C>G, C>T, T>A, T>C, T>G). The immediate 5' and 3' sequence contexts for each mutation was extracted from the CanFam3.1 dog reference genome (Lindblad-Toh et al., 2005) yielding 96 mutation types.

3.10.5.1 Signature fitting

We performed mutational signature fitting in order to estimate the number of mutations contributed by different exposures to the CTVT mutational spectrum. Validated mutational signatures were obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC) database (Forbes et al., 2015) and renormalised to the CanFam3.1 dog reference genome (Lindblad-Toh et al., 2005). In addition, we generated a "dog germline" signature, from the germline mutational spectrum of a Greenland sledge dog (see Section 3.10.4.2).

Murchison et al. (2014) previously showed that COSMIC mutational signature 1 (5-methylcytosine deamination), signature 5 (unknown etiology), and signature 7 (ultraviolet light exposure) are operative in CTVT, and that these three signatures are sufficient to describe the pattern of somatic substitutions observed in CTVT. We therefore fitted these three signatures, together with the dog germline signature, to the CTVT tumour-only mutation spectrum (Table 3.7). Results were similar to previous findings (Murchison et al., 2014). Signatures were fitted using the sigfit v1.1.0 package in R (Gori and Baez-Ortega, 2018). Simulations were run using 100 chains with 10,000 iterations each. Importantly, the Dog Germline signature accounted for only 5.5% of the tumour-only SNVs, suggesting that the majority of the SNVs in this set are indeed somatic.

3.10.6 Copy number analysis

Average mappability and GC-content were generated for the dog reference genome (Lindblad-Toh et al., 2005; http://www.ensembl.org/Canis_familiaris/Info/Index) with the generateMap, mapCounter, and gcCounter tools in the HMMcopy package (Ha et al., 2012). GC content and genomic mappability biases for read counts in non-overlapping 1 kb windows were corrected using HMMcopy. Copy number estimation was then performed on GC- and mappability-corrected read counts using a bespoke copy number calling pipeline (https://github.com/ymk1/cnv_pipeline.git).

Tumour purity in 608T and 609T was evaluated based on the VAF distribution of tumour only SNVs. Tumour purity was estimated as follows:

$$Purity = 2 * VAF_{med}$$

Where VAF_{med} is the median VAF value of tumour-only SNVs. Using this method, 608T was estimated to be 49.3% CTVT cells and 609T was estimated to be 65% CTVT cells.

3.10.6.1 Clonal tumour-unique SNVs

We categorised tumour-unique SNVs in 608T and 609T as either clonal or subclonal, that is, present in all or a fraction of tumour cells within a sample, respectively (see Table 3.6). To do this, we first examined the VAF distributions of germline SNPs in 608T and 609T for each copy number (CN) state (CN1, CN2, CN3, CN4, CN6). We used a Gaussian mixture model (k = 2), implemented using the R package MCLUST (Fraley et al., 2012), to define VAF clusters for heterozygous and homozygous SNPs. Next, we fitted this model to VAF distributions of tumour-unique SNVs. SNVs that fell below the 5% lower bound were defined as subclonal; all other SNVs were considered clonal. The results of this analysis are shown in Table 3.8.

3.10.7 Phenotype information

Considering the evidence of introgression between wild North American canids and the pre-contact domestic dog population, we assessed the presence of a marker associated with melanism, which has introgressed from dogs into North American gray wolves and coyotes, in the higher coverage pre-contact dog genomes (Port du Choix sample: AL3194, ~2×; Weyanoke old town sample: AL3223, ~0.5×).

3.11 Results

In the following subsections, I present results that follow from the research questions posed at the beginning of this chapter (Section 3.3.4).

3.11.1 Ancestry of the CTVT founder

To investigate the ancestry of the CTVT founder dog, we analysed two previously published CTVT genomes (Murchison et al., 2014) alongside a panel of modern and ancient canid genomes. We generated low coverage nuclear genomes of seven pre-contact dogs (~0.005 - 2.0×), including the earliest dog remains on the continent (Perri et al., 2018), from six locations in North America with time frames spanning

609T (Mother)	609T (Son)	Interpretation
Clonal	Clonal	Occurred before Mother-Son MRCA cell
Clonal	Subclonal	Son infection was polyclonal with clone a and clone b; clones a and b were sampled in the son, only clone a was sampled in the mother
Clonal	Absent	Arose in $MRCA_{609T}$ cell; $MRCA_{609T}$ cell was either unique to mother or we did not sample that subclone in the son
Subclonal	Clonal	We sampled a fraction of tumour in son that derived from a subclone in the mother's tumour.
Subclonal	Subclonal	Polyclonal seeding and sampling: both subclones exist in son and in mother
Subclonal	Absent	Occurred subsequent to $MRCA_{609T}$, was not transmitted to son or we did not sample it in son
Absent	Clonal	Occurred in $MRCA_{608T}$; either arose in son subclone, or was present in mother but we sampled a different subclone in the mother
Absent	Subclonal	Occurred subsequent to $MRCA_{608T}$; either arose in son subclone, or was present in 609T and there was polyclonal seeding but we sampled a different subclone in the mother

Table 3.6 Interpretation of clonal and subclonal mutations in mother-son direct transmission pair

approximately 9,000 years. These data allowed us to assess the relationship between CTVT and ancient dogs as well as where native dogs in the Americas originated.

A neighbour-joining tree analysis generated using single nucleotide polymorphisms confirmed that PCD individuals clustered in a distinct monophyletic lineage that is more closely related to global dog populations than to either Eurasian or North American wolves (Fig. 3.4b). The closest related sister clade to PCD consisted of present-day Arctic dogs from the Americas (including Alaskan malamutes, Greenland dogs and Alaskan huskies) and Eurasia (Siberian huskies). TreeMix, *D*-statistics and outgroup f_3 statistics results were concordant with this phylogenetic topology. Importantly, CTVT clustered with PCDs on neighbour-joining trees (bootstrap=100), a Bayesian tree (posterior probability [PP] = 1), TreeMix, PCA and admixture graphs (Fig. 3.4). This result is further supported by both outgroup f_3 statistics and *D*-statistics (Fig. 3.10). Altogether, these findings indicate that the CTVT founder dog is more closely related to PCDs than to modern Arctic dogs.

3.11.2 The origin of pre-contact dogs

We generated 71 complete mitochondrial genomes from archaeological dog remains in North America and Siberia (Fig. 3.1a; Tables 3.1 and 3.2). We analysed these alongside 145 mitogenomes derived from ancient dogs and a global population of modern canids. Phylogenetic trees constructed from these sequences confirmed that all sampled pre-contact dogs, apart from one individual, form a highly supported monophyletic clade within dog haplogroup A (bootstrap value=87; Fig. 3.5). The most closely related mitochondrial lineage to the PCD clade, from the sampled dog populations, were ancient sledge dogs from Zhokhov island in Eastern Siberia (Pitulko and Kasparov, 2017).

Molecular clock analyses (Fig. 3.6) based on mtDNA suggest that all pre-contact dogs shared a matrilineal common ancestor ~14,666 years ago (95% highest posterior density interval (HPDI): 12,965-16,484 years ago), that had diverged from an ancestor shared with Zhokhov dogs roughly 1000 years earlier, ~15,606 years ago (95% HPDI:13,739-17,646 years ago). These time frames are broadly concurrent with early migrations into the Americas (Jakobsson et al., 2017; Moreno-Mayar et al., 2018; Raghavan et al., 2015).

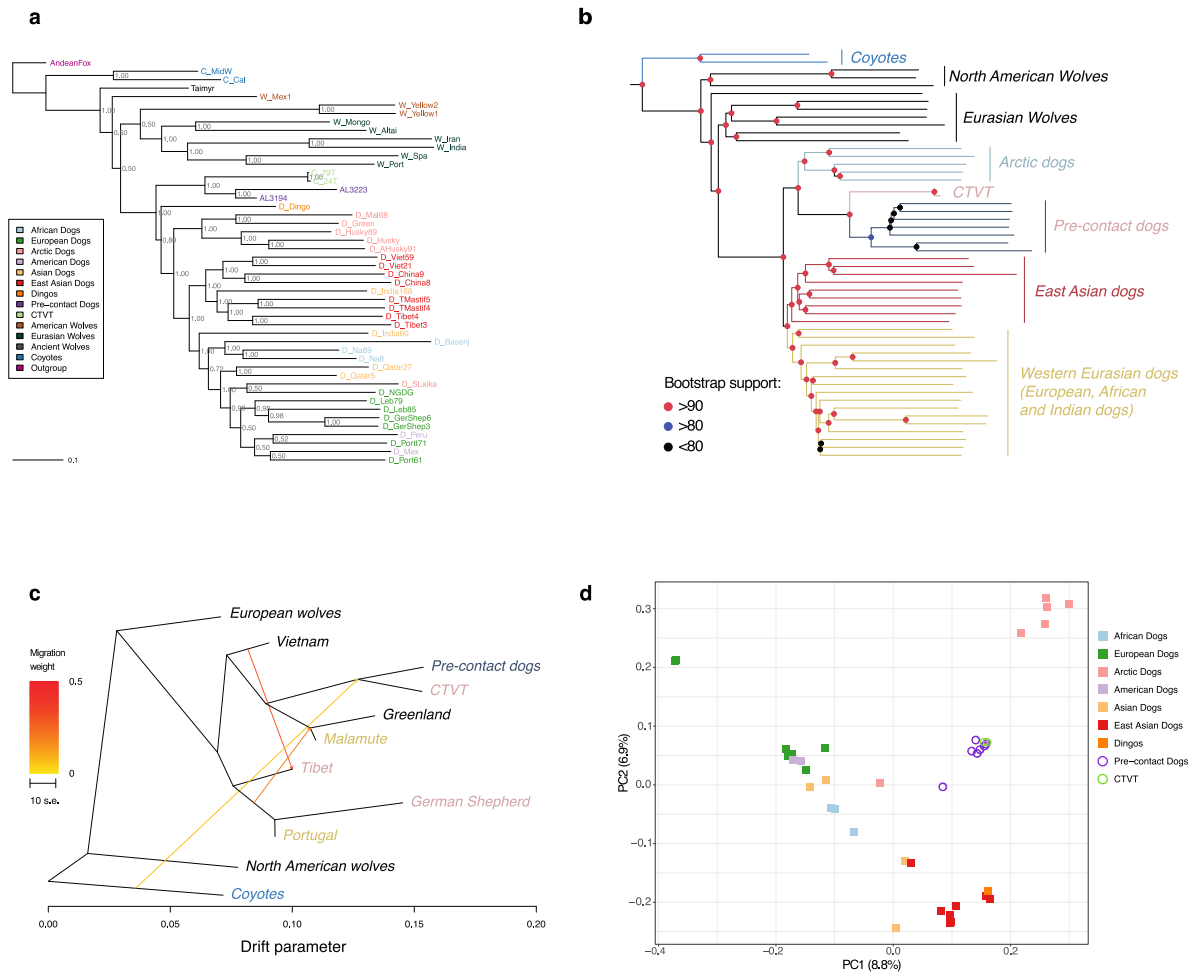


Fig. 3.4. a. Bayesian tree based on ~26K transversions. **b.** A neighbour-joining tree built with whole genomes. **c.** An admixture graph constructed with Treemix (on the basis of transversions) (Pickrell and Pritchard, 2012) depicting the relationship between PCDs (including the Port au Choix (AL3194) and Weyanoke Old Town (AL3223) samples) and other dog, wolf, and CTVT populations. The scale bar shows 10 times the average standard error (s.e.) of the entries in the sample covariance matrix (Pickrell and Pritchard, 2012). **d.** Principal Component Analysis plot (PC1 versus PC2) of 44 dog samples (excluding wolves and coyotes) based on 2,063,129 SNPs ascertained using the genome-wide data set. All pre-contact dog and CTVT samples were projected. Figure adapted from Ní Leathlobhair et al. (2018).

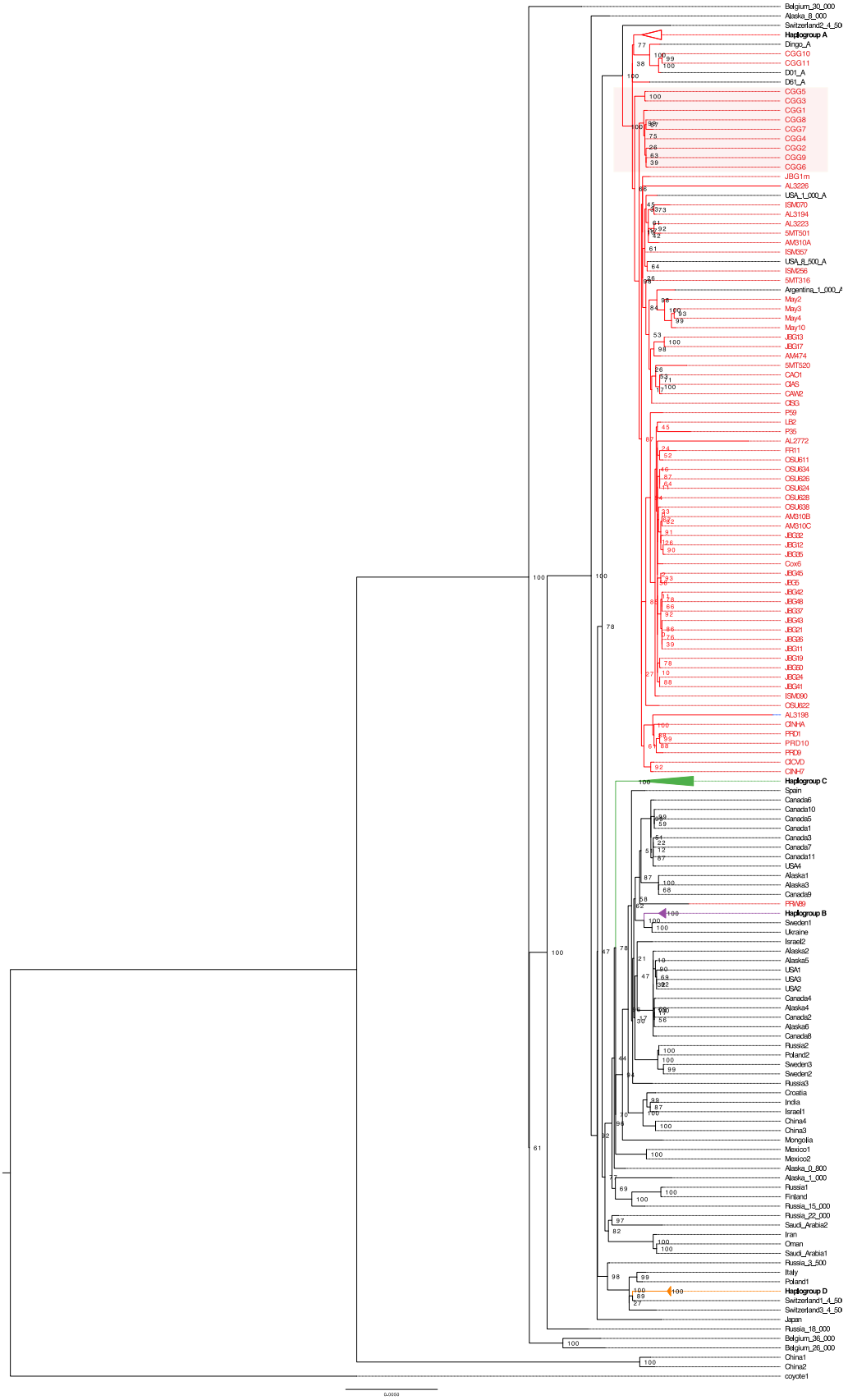


Fig. 3.5. Maximum likelihood tree based on mtDNA data. The four major dog haplogroups are indicated: A (red), B (purple), C (green), D (orange). Red tip labels are newly sequenced samples. Light red highlighted clades (CGG1-9) are Zhokhov Island samples (~9 kya sled dogs). CGG10-11 are more recent sledge dogs from Siberia (~1.5 kya). Node labels indicate bootstrap replicates.

While it was found that the major dog mitochondrial haplogroups (A, B, C, D) can be distinguished using the control region (Fig. 3.7), it is not possible to distinguish between pre-contact dogs and other dogs within haplogroup A using the mtDNA control region alone.

It should be noted that samples CGG10 and CGG11 fall outside of the Zhokhov/pre-contact clade (Fig. 3.5) as these are relatively recent sledge dogs from Chukotka, Russia (~1.5 kya; Table 3.1). Lastly, one sample from the Prince Rupert Harbour site (PRW89; ~1.5 kya) clustered with North American wolves (Fig. 3.5). Wolves and dogs were poorly distinguished at this site (Section 3.4.1.1); it is possible that this sample derives from a wolf or possibly represents a wolf-dog hybrid.

As demonstrated in Chapter 2, the CTVT founder dog's mitogenome has been replaced as a result of multiple mitogenome horizontal transfers from dog hosts to CTVT (Rebbeck et al., 2011; Strakova et al., 2016); thus, we could not determine the mitochondrial haplotype of the CTVT founder dog, and analysis of the CTVT founder's ancestry was limited to the nuclear genome. However, the combined mitochondrial and nuclear results confirm that CTVT is more closely related to PCD than to any other dog populations.

These analyses also indicate that PCD were not domesticated *in situ* from North American wolves but instead migrated alongside humans into the Americas via Beringia from a population that was closely related to modern Arctic dogs.

3.11.3 Estimating the CTVT mutation rate and time-of-origin

Presented with the finding that CTVT was closely related to an early lineage of dogs present in Siberia and the Americas, resolving the 'age' of CTVT became crucial to understanding whether CTVT arose before or after dogs entered North America. In order to gain a more accurate estimate of the CTVT time-of-origin, we sequenced the genomes of two CTVT tumours involved in a naturally occurring direct transmission event, 609T (mother) and 608T (pup). This section describes the results of the somatic variation analysis described in Section 3.10 which were used to derive a lower bound for the somatic mutation rate in CTVT.

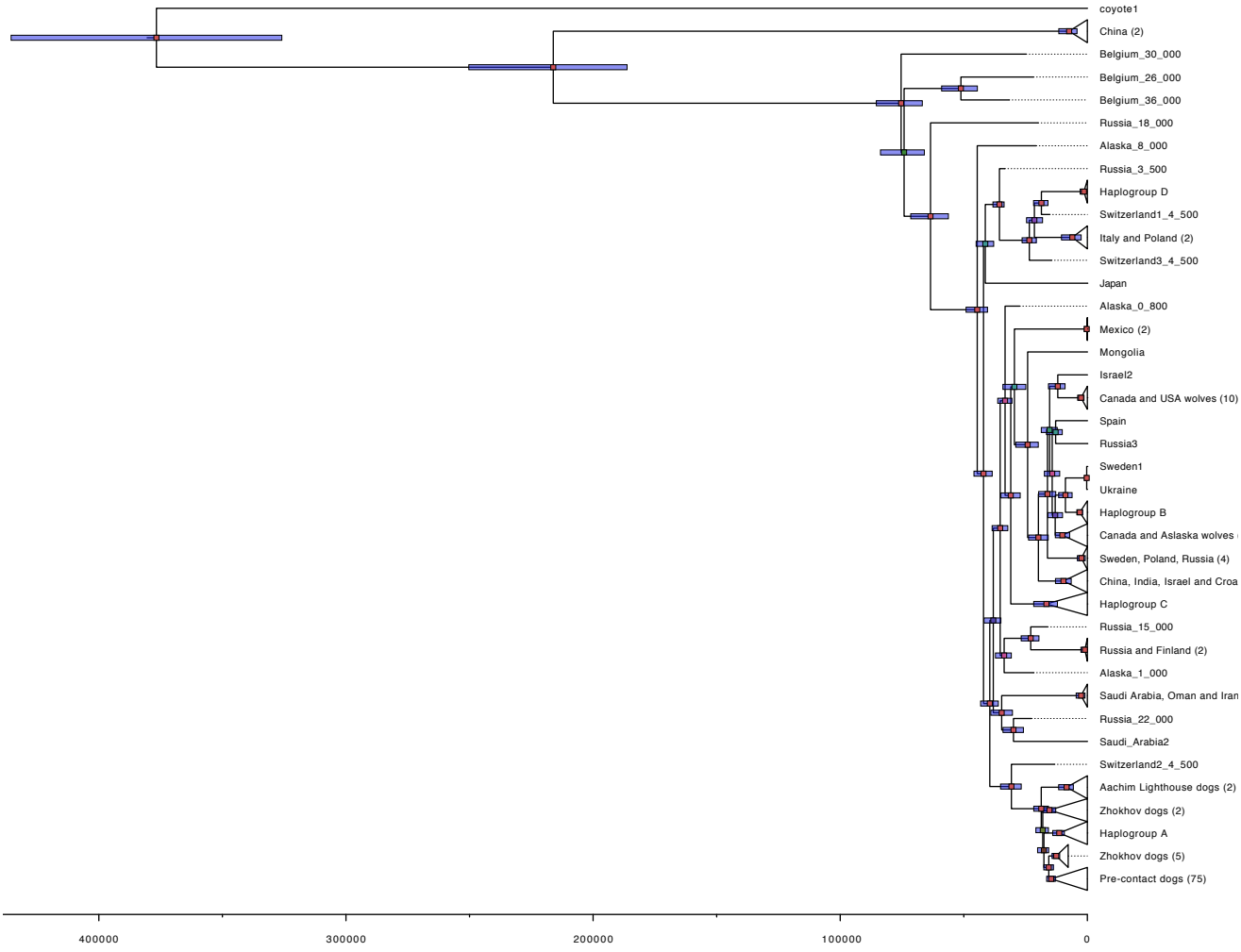


Fig. 3.6. Bayesian tree of wolf, coyote and dog mtDNA, including modern and pre-contact dog samples. Red, purple and green circles represent nodes with >0.9, >0.7 and >0.5 posterior probability respectively. Blue bars represent confidence intervals of divergence times (scaled in years before present). Figure adapted from Ní Leathlobhair et al. (2018).

Fig. 3.7. Maximum likelihood tree based on 605 bp of the control region. All samples starting with prefix HQ were obtained from Leonard et al. (2002) while all samples starting with prefix AY were obtained from van Asch et al. (2013). The four major dog haplogroups are indicated with different branch colours: A (red), B (purple), C (green), D (orange). All pre-contact dogs are highlighted in red. Node labels indicate bootstrap replicates.

Table 3.7 Fitting of COSMIC mutational signatures 1, 5, 7 and Dog Germline signature to 1,933,897 CTVT tumour-only SNVs shared between 608T and 609T.

Mutational signature	Fit	Number of SNVs
1	Mean (15.4%)	297,820.1
	Lower bound (15.2%)	293,952.3
	Upper bound (15.5%)	299,754
5	Mean (40.1%)	775,492.7
	Lower bound (40.0%)	773,558.8
	Upper bound (40.3%)	779,360.5
7	Mean (38.95%)	753,252.9
	Lower bound (38.9%)	752,285.9
	Upper bound (39%)	754,219.8
Germline	Mean (5.5%)	106,364.3
	Lower bound (5.3%)	102,496.5
	Upper bound (5.7%)	110,232

3.11.3.1 CTVT mutation rate

The clocklike mutational signature 1 (Alexandrov et al., 2015) is largely composed of cytosine-to-thymine (C>T) mutations at NpCpG dinucleotides, where N is any base (N[C>T]G mutations). Of the 1,933,897 CTVT tumour-only SNVs (Section 3.10.4.2), shared by 608T and 609T, 222,072 were N[C>T]G (Fig. 3.8).

We identified 183 and 174 clonal mutations unique to tumours 608T (pup) and 609T (mother), respectively (Table 3.8). It was determined that 27 and 23 clonal tumour-unique N[C>T]G SNVs had arisen in 608T and 609T, respectively, since divergence from their most recent common ancestor (MRCA).

In order to estimate the CTVT mutation rate, we needed to estimate the time intervals during which clonal tumour-unique mutations arose in 608T and 609T. These time intervals (i_{608T} and i_{609T}) correspond to:

$$i_{608T} = t_{MRCA\text{-}608T} - t_{MRCA\text{-}608T/609T}$$
$$i_{609T} = t_{MRCA\text{-}609T} - t_{MRCA\text{-}608T/609T}$$

where $t_{MRCA\text{-}608T}$ and $t_{MRCA\text{-}609T}$ are time-points defining the MRCA cells of 608T and 609T respectively, and $t_{MRCA\text{-}608T/609T}$ is the time-point defining the MRCA cell of both

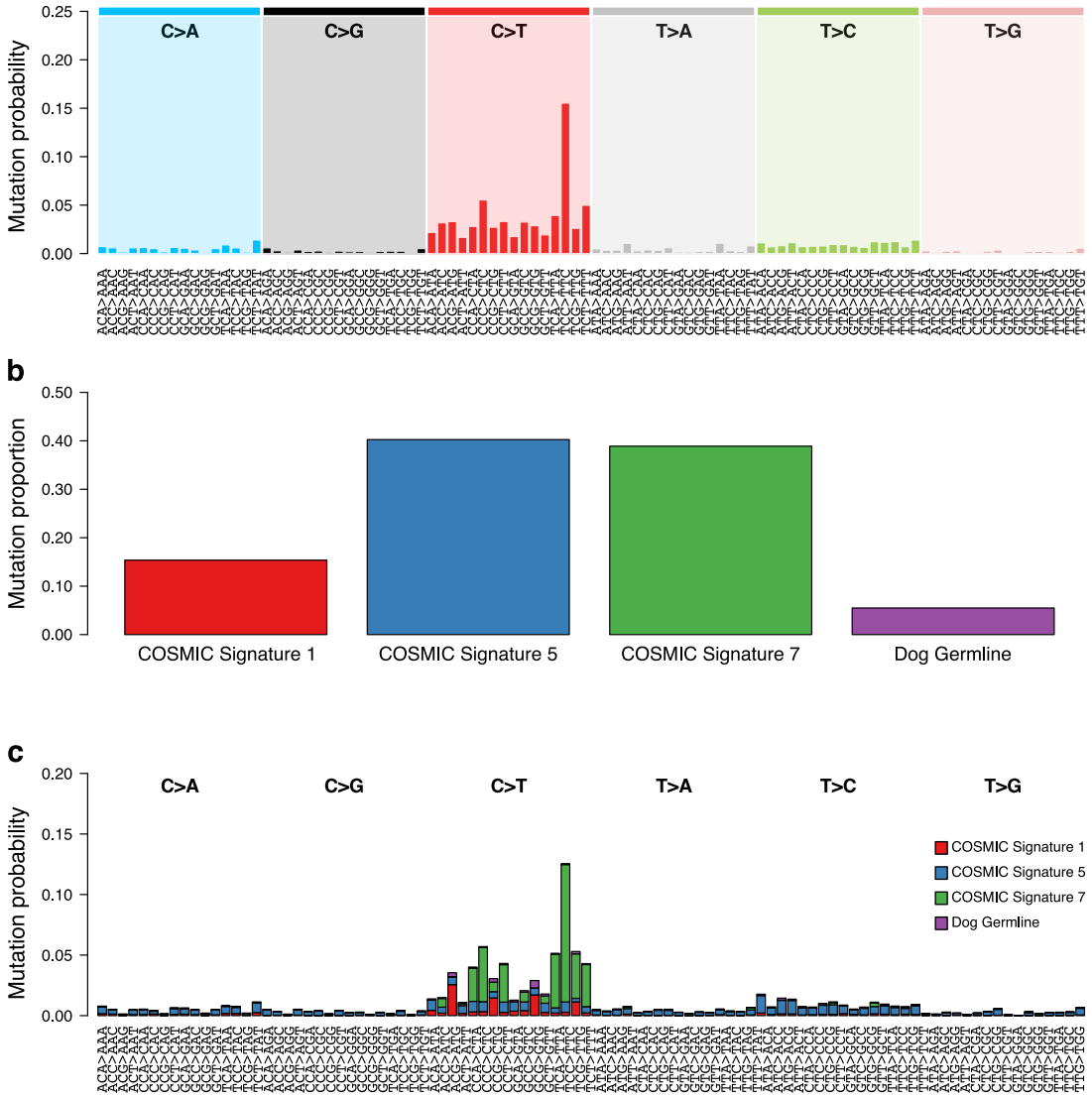


Fig. 3.8. a. CTVT mutation spectrum. 1,933,897 tumour-only mutations in CTVT are displayed by mutation type (in pyrimidine context) with immediate 5′ and 3′ context. Each of the 96 mutation classes is displayed on the horizontal axis. Mutation proportions are displayed relative to CanFam3.1. **b.** Fraction of CTVT tumour-only mutations attributable to COSMIC Signatures 1, 5, 7, and the dog germline signature, estimated using sigfit. **c.** Reconstruction of CTVT tumour-only spectrum using COSMIC signatures 1,5 and 7 and the dog germline signature. Figure adapted from Ní Leathlobhair et al. (2018).

Sample	CN state	Total	Clonal
608T	CN1	40	33
	CN2	140	125
	CN3	22	21
	CN4	3	3
	CN6	1	1
		206	183
609T	CN1	43	30
	CN2	159	125
	CN3	21	15
	CN4	5	4
		228	174

Table 3.8 Number of clonal tumour-unique SNVs in 608T and 609T
CN, copy number. "Total" refers to the complete set of tumour-unique SNVs, "Clonal", indicates only those that are clonal.

608T and 609T.

We assumed that the clones that seeded the sampled 608T and 609T biopsies diverged in the period after infection of the mother (Dog 609) but before transmission to the pup (Dog 608) i.e. that $t_{MRCA\text{-}608T/609T}$ occurred during 609T tumour development. This implied that the earliest time point for $t_{MRCA\text{-}608T/609T}$ would coincide with the time at which Dog 609 (mother) was infected with CTVT, i.e. month 0.

We assumed that the mother (Dog 609) was infected during the heat cycle in which she conceived Dog 608. While we cannot be certain that this assumption is valid, we observed that the mother's tumour appeared to be of a similar size to the pup's tumour. Unless CTVT tumours have a large variation in growth rate, we believe it is unlikely that the mother was infected with CTVT in the heat cycle prior to that in which she conceived Dog 608.

Given that the gestation period in domestic dogs can range from 57-72 days (Concannon et al., 1983; Gavrilovic et al., 2008; Okkens et al., 1993), we estimated that Dog 609 was infected with CTVT approximately 2 months prior to when her son, Dog 608, was born and infected. This implies that the latest time-point for $t_{MRCA\text{-}608T/609T}$ is 2 months after Dog 609 was infected with CTVT. Thus $t_{MRCA\text{-}608T/609T}$ = 0 to 2 months,

where month 0 as the month at which the mother (Dog 609) was infected with CTVT.

Assuming no polyclonal seeding, $t_{MRCA\text{-}608T}$ either arose in the pup (Dog 608), or was the cell that transmitted from mother (Dog 609) to pup (Dog 608), and thus $t_{MRCA\text{-}608T}$ = 2 to 12 months, where month 0 is the month at which the mother (Dog 609) was infected with CTVT, and 12 months corresponds to the time of sampling.

$t_{MRCA\text{-}609T}$ could have occurred at any time during 609T tumour development. Thus $t_{MRCA\text{-}609T}$ = 0 to 12 months, where month 0 is the month at which Dog 609 (mother) was infected with CTVT, and 12 months corresponds to the time of sampling. Thus, we estimate both i_{608T} and i_{609T} to be 0 to 12 months.

We determined that 608T and 609T had acquired 183 and 174 mutations since their divergence from their MRCA ($MRCA_{608T/609T}$) and before the MRCA of the clone biopsied in 608T ($MRCA_{608T}$) and the MRCA and the clone biopsied in 609T ($MRCA_{609T}$). These mutations will likely have arisen as part of clocklike ageing-associated mutational signatures 1 and 5 and possibly as part of mutational signature 7 (exposure to ultraviolet light).

The signature 1 mutation rate is believed to be highly dependent on cell division (Alexandrov et al., 2015). Due to the small number of tumour-unique mutations, signature fitting cannot give us an accurate estimate of the respective contributions of these signatures to tumour-unique SNV sets. As mutational signature 1 is largely composed of N[C>T]G mutations (where N is any base), we used N[C>T]G as a proxy for signature 1. 27 and 23 N[C>T]G mutations were unique to 608T and 609T respectively, and 222,072 N[C>T]G mutations arose in the somatic lineage from the CTVT founder dog until $MRCA_{608T\text{-}609T}$.

As 608T harbours more clonal N[C>T]G mutations than 609T (Table 3.8), we infer that $MRCA_{608T}$ existed more recently than $MRCA_{609T}$. Assuming that i_{608T} is up to 12 months (see above), then the slowest rate at which N[C>T]G mutations accumulate is 27 N[C>T]G mutations/ year.

Applying this rate to the complete set of clocklike somatic mutations in the CTVT lineage (221,385 N[C>T]G mutations, see above), we estimated that the CTVT founder

dog lived up to 8,225 years ago. This time frame postdates the initial arrival of dogs into the Americas, and is compatible with the possibility that CTVT may have originated in a dog living in North America.

3.11.3.2 Comparison with mutation rates in human cancer

Signature 1 mutation rate varies between human cancer tissue types (Alexandrov et al., 2015). The mutation rate lower bound that we have derived for CTVT N[C>T]G mutations (>12.56 mutations / Gb / year) is comparable to the N[C>T]G mutation rates found in human cancers (Alexandrov et al., 2015). Cervical cancer was reported to have the highest estimated rate of accumulation of N[C>T]G somatic mutations of 36 human cancer types (16.61 N[C>T]G somatic mutations / Gb / year)(Alexandrov et al., 2015). Applying the cervical cancer N[C>T]G mutation rate to the CTVT lineage would provide an estimate of 6,195 years since CTVT's origin.

3.11.4 Pre-contact dogs and global dog clade structures

Previous studies of nuclear genome data have highighted two major modern clades of dogs worldwide: a West Eurasian clade (including European, Indian, and African dogs) and an East Asian clade (including dingoes) (Frantz et al., 2016; van Asch et al., 2013; Wang et al., 2016). These analyses placed modern Arctic dogs such as Siberian huskies and Greenland sledge dogs in either West Eurasian (Vonholdt et al., 2010) or East Asian (Frantz et al., 2016) groups. Phylogenetic analyses of nuclear genomes indicate that the clade formed by PCD and modern Arctic dogs (PCD/Arctic) is basal to both West Eurasian and East Asian dogs and suggests the existence of a third monophyletic clade of dogs (Fig. 3.4b). Outgroup f_3 statistics also support the observation that PCD/CTVT and Arctic breeds are equally related to all other dogs, except for Basenji and one Indian dog, which could be due to admixture from wolves into these two samples (e.g. Indian wolf or golden wolf; Fig. 3.10). Arctic breeds also appear more closely related to PCD/CTVT than any other dog population (Fig. 3.10b and c).

Although nodes leading to all three clades were well supported, the relationships between these groups were inconsistent across different analyses. Admixture analysis showed that all East Asian dogs, excluding Vietnamese village dogs, are significantly admixed with European dog populations (Table 3.9). Such disproportionate ad-

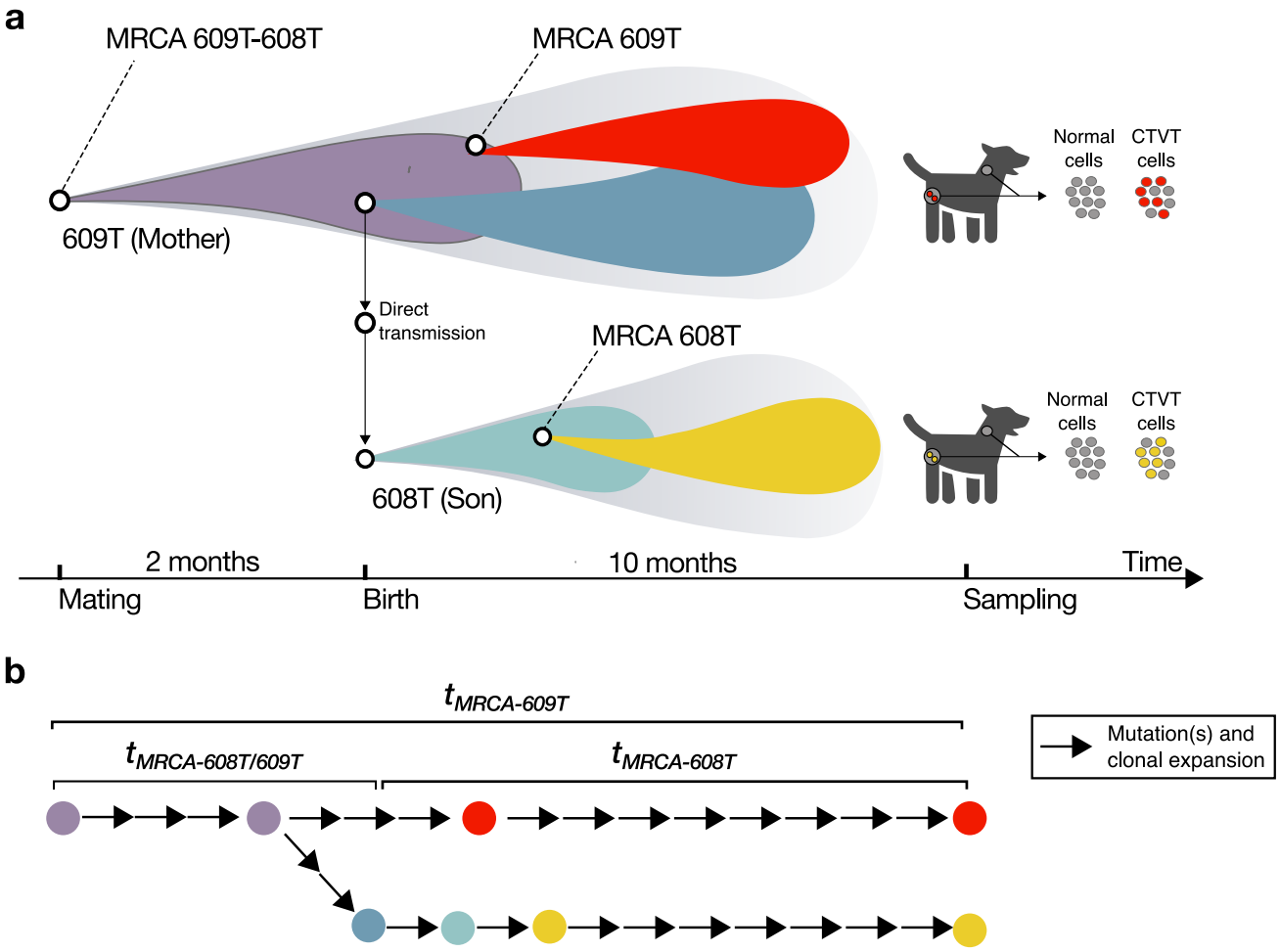


Fig. 3.9. Tumour evolution in 609T and 608T
a.Clonal expansion occurs in association with the acquisition of de novo mutations. MRCA, most recent common ancestor. With time, the cells (MRCA 609T and MRCA 608T) that give rise to the sampled CTVT tumours appear. **b.** Schema of the interval estimates used in estimating the CTVT mutation rate. A founder cell from within the mother's lesion is transmitted to the pup, initiating a new tumour growth (end of $t_{MRCA-608T/609T}$ and beginning of $t_{MRCA-608T}$); time of birth acts as a hard upper bound on $t_{MRCA-608T}$ and $t_{MRCA-609T}$. t_{MRCA} is the time of the MRCA of the tumour cell.

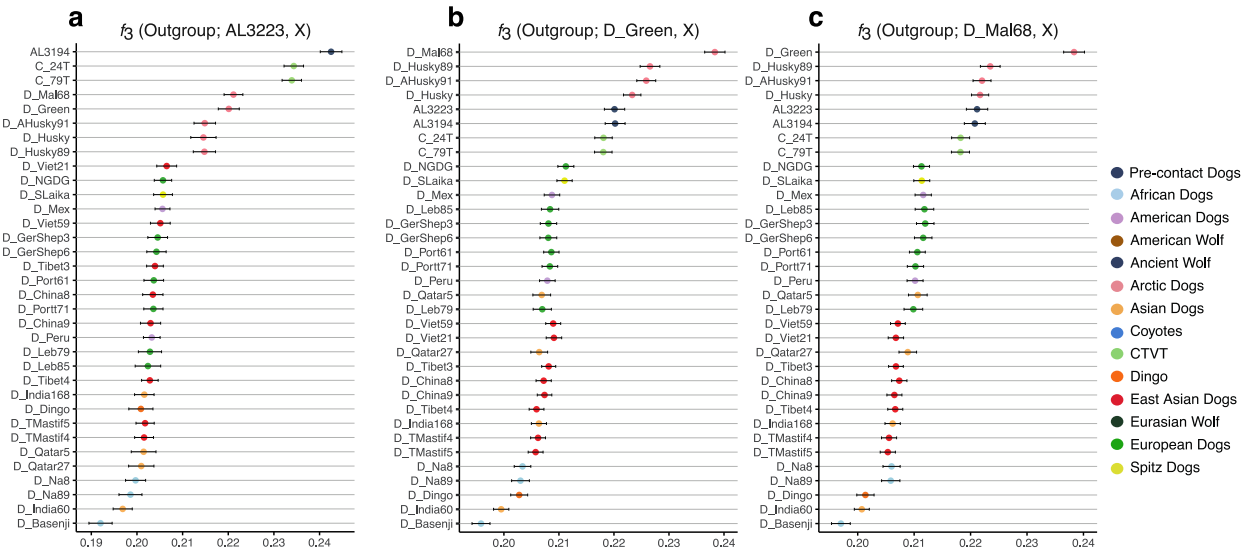


Fig. 3.10. Shared genetic drift measured by f_3(Outgroup; Y, X) where Y is either **a**. Weyanoke Old town dog (AL3223), **b**. Alaskan Malamute (D_Mal68), **c**.Greenland sledge dog (D_Green) and X represents modern dog populations. Error bars represent 1 standard error (SE). Figure adapted from Ní Leathlobhair et al. (2018).

mixture could affect the topology of the tree. For example, if specific East Asian dogs where there is evidence of gene flow with European dogs (Wang et al., 2016) were excluded, East Asian dogs became the most basal clade, and PCD and modern Arctic dogs appeared as a sister clade to West Eurasian dogs. Conversely, admixture graphs and TreeMix suggested that West Eurasian dogs were more basal and that PCD/Arctic dogs were closest to East Asian dogs. These conflicting topologies, illustrated by previous analyses of nuclear data (Frantz et al., 2016; Larson et al., 2012; Wang et al., 2016), could result either from a near simultaneous divergence of all three lineages or from substantial post-divergence gene flow among Eurasian dogs, as previously suggested (Frantz et al., 2016; Wang et al., 2016).

3.11.5 Pre-contact dog ancestry in modern Arctic dogs

As we have previously outlined (Section 3.11.1), nuclear genome data indicated that modern Arctic dogs sampled from both Siberia and North America cluster in a distinct phylogenetic group that forms a sister taxon to pre-contact dogs (Fig. 3.4b) suggesting that PCDs are not the direct ancestor of modern American Arctic (Alaskan

Table 3.9 *D*-statistics for *D*(Outgroup, Portuguese village dogs, Vietnamese village dogs, Asian dogs). The Andean fox was used as an outgroup for these analyses. *n* corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks. Statistics show evidence of admixture in all East Asian populations tested in this study.

Pop 2	Pop 3	Pop 4	*D*	*Z*	BABA	ABBA	n
D_Port61	D_Viet21	D_China8	0.1029	11.106	52,952	43,074	934,713
D_Port61	D_Viet21	D_China9	0.0557	6.079	48,963	43,796	924,557
D_Port61	D_Viet59	D_China8	0.0965	10.469	52,395	43,172	941,437
D_Port61	D_Viet59	D_China9	0.0481	5.589	47,976	43,576	920,022
D_Port61	D_Viet21	D_Tibet3	0.0663	6.823	49,504	43,352	922,535
D_Port61	D_Viet21	D_Tibet4	0.0599	6.610	49,756	44,134	914,979
D_Port61	D_Viet59	D_Tibet3	0.0575	6.277	49,026	43,693	929,474
D_Port61	D_Viet59	D_Tibet4	0.0513	5.871	49,127	44,333	921,845
D_Port61	D_Viet21	D_Mastif4	0.0714	7.278	49,690	43,069	894,260
D_Port61	D_Viet21	D_Mastif5	0.0643	7.229	49,757	43,747	913,367
D_Port61	D_Viet59	D_Mastif4	0.0639	6.791	48,007	42,240	898,585
D_Port61	D_Viet59	D_Mastif5	0.0587	6.929	48,421	43,054	917,851

malamutes, Alaskan huskies, and Greenland dogs) and Eurasian Arctic (Siberian huskies) dogs. It is possible that modern Arctic dogs are the descendants of dogs introduced to the continent by the Paleo-Eskimos (~6,000 years ago) or by the Thule people (~1,000 years ago) (Raghavan et al., 2014).

Moreover, both mitogenomic and low-coverage nuclear genomic data from a late Paleo-Eskimo dog from Alaska (Uyak site sample AL3198; Fig. 3.1) indicate that this dog was more closely related to PCDs than to modern American Arctic dogs. This finding suggests that modern American Arctic dogs are not the descendants of Paleo-Eskimo dogs and that Paleo-Eskimos brought dogs from Siberia derived from the same source population as PCDs or adopted local dogs in North America.

We also used *D*-statistics on genome-wide data to test for admixture from PCD into Arctic breeds since their MRCA. We found that both Alaskan malamute and Greenland sledge dogs have a significant amount of ancestry from PCD (Table 3.10). We tested whether this signal could be due to admixture from Eurasian dogs into Siberian husky dogs (making derived alleles in Alaskan malamute and Greenland

Table 3.10 *D*-statistics for *D*(Outgroup, PCD (AL3194), Arctic dogs, Arctic dogs). *n* corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks. Statistics with *Z*>3 and *Z*<-3 are shown in bold.

Pop 3	Pop 4	*D*	*Z*	ABBA	BABA	n
D_Mal68	D_Green	0.0001	0.010	38,564	38,555	926,919
D_Husky89	D_AHusky91	0.0059	0.416	29,524	29,174	749,696
D_Husky	D_Husky89	0.0166	1.254	33,365	32,275	797,067
D_Husky	D_AHusky91	0.0176	1.324	37,461	36,164	913,221
D_Husky89	D_Green	0.0437	3.261	38,311	35,101	790,543
D_Husky89	**D_Mal68**	**0.0455**	**3.384**	**39,192**	**35,781**	**775,433**
D_AHusky91	**D_Green**	**0.0470**	**3.555**	**44,116**	**40,158**	**897,276**
D_AHusky91	**D_Mal68**	**0.0450**	**3.596**	**44,737**	**40,883**	**884,940**
D_Husky	**D_Green**	**0.0619**	**5.021**	**48,948**	**43,246**	**955,940**
D_Husky	**D_Mal68**	**0.0609**	**5.050**	**49,177**	**43,528**	**942,665**

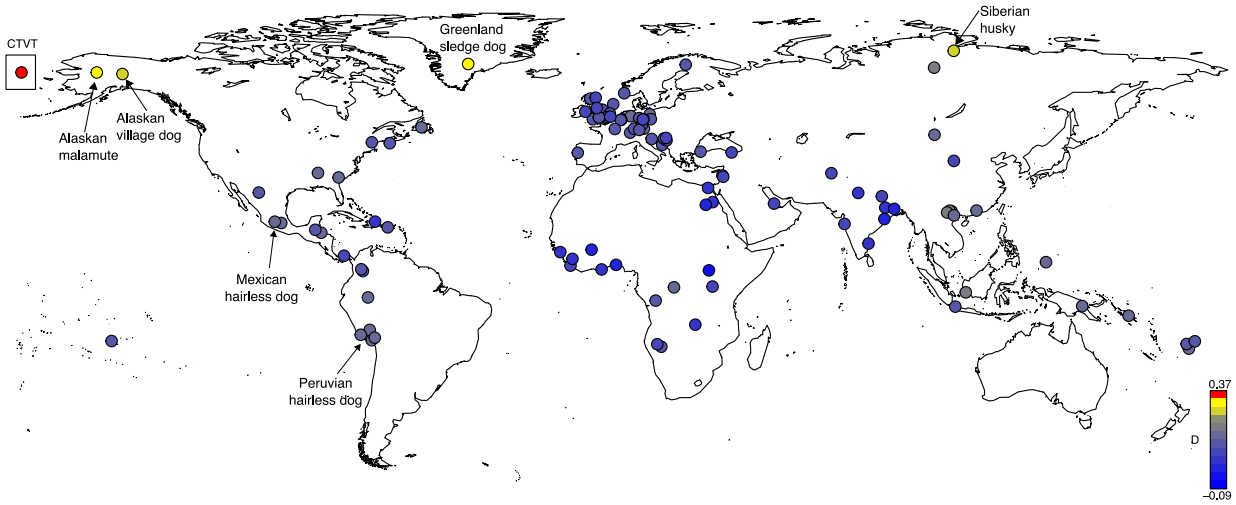


Fig. 3.11. A map showing the locations of dog populations obtained from Shannon et al. (2015) and their degrees of relatedness (*D*-statistics) to the ~4,000 year old Port au Choix dog (AL3194). Higher values (in red) represent closer relatedness. The location of the founder CTVT individual, labelled in the plot, is unknown. Figure adapted from Ní Leathlobhair et al. (2018).

sledge dogs (GSD) match PCD more often; Table 3.12) as *D*-statistics suggest that all sampled modern Arctic breeds are a mixture of the basal lineage (leading to the CTVT founder and PCD) and of the Eurasian dog lineage (Table 3.11). We found evidence that the Siberian husky and Alaskan malamute genomes analysed in this chapter received gene flow from European dogs (Table 3.12). However, we found no evidence that GSD received gene-flow since the MRCA of Arctic breeds, suggesting that Eurasian admixture did not affect our result. This could also be the result of ancient substructure within Arctic dogs.

We found no signal that either Alaskan malamute or Greenland sledge dogs shared an excess of derived alleles with PCD compared with each other (*D*(Outgroup, AL3194, Alaskan malamute, Greenland sledge dog) = -0.0001, sd = -0.010). *D*-statistics suggest that there was no additional gene-flow between PCD and these dog lineages since their divergence from each other. This finding suggests that the signal detected above (excess shared derived alleles between American Arctic dogs and PCD) is due to ancient substructure within Arctic dogs (Eriksson and Manica, 2012).

3.11.6 Pre-contact ancestry in modern American dogs

While our results so far have indicated that only modern Arctic dogs and CTVT share a genetic affinity with PCD, multiple studies have argued that modern American dog populations possess a genetic signature from indigenous dogs (Parker et al., 2017; Shannon et al., 2015; van Asch et al., 2013).

To test this hypothesis, we analysed nuclear data obtained from more than 5,000 modern dogs (including American village dogs) genotyped on a semi-custom 185K SNP array (Shannon et al., 2015). The results of the f_4 test depicted in Fig. 3.2 show that, apart from Alaskan Village dogs, we detect no significant signal of pre-contact ancestry in modern North American populations (α always <4% and Z always < 3; Table 3.13). Alaskan village dogs, on the other hand, have between 11-20% (Z always > 4.5; Table 3.13) ancestry derived from pre-contact dogs. f_4 ratio analysis did not detect a significant admixture signal from PCDs in any modern American dogs of European ancestry (Table 3.5 for list of American dogs).

Alaskan village dogs were the population of North American village dog with the

Table 3.11 *D*-statistics for *D*(Outgroup, Asian or European dogs, PCD or CTVT, Arctic dogs). *n* corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks.

Pop2	Pop3	Pop4	*D*	*Z*	BABA	ABBA	n
D_China8	AL3194	D_AHusky91	0.0752	7.65	47,384	40,759	879,166
D_China8	AL3194	D_Green	0.0606	6.01	47,406	41,994	920,449
D_China8	AL3194	D_Husky	0.0693	7.253	51,412	44,755	936,527
D_China8	AL3194	D_Husky89	0.0825	8.407	41,900	35,513	768,218
D_China8	AL3194	D_Mal68	0.0655	6.538	46,698	40,958	907,759
D_China8	C_24T	D_AHusky91	0.0907	9.461	61,440	51,230	1,115,093
D_China8	C_24T	D_Green	0.0753	7.523	61,012	52,472	1,167,919
D_China8	C_24T	D_Husky	0.0819	8.781	66,255	56,230	1,189,650
D_China8	C_24T	D_Husky89	0.0987	9.798	54,693	44,866	985,295
D_China8	C_24T	D_Mal68	0.0796	8.128	60,682	51,739	1,151,059
D_China9	AL3194	D_AHusky91	0.064	6.644	45,860	40,348	849971
D_China9	AL3194	D_Green	0.0812	8.035	47,827	40,641	893422
D_China9	AL3194	D_Husky	0.0684	7.323	50,481	44,022	905549
D_China9	AL3194	D_Husky89	0.087	8.783	41,939	35,228	760651
D_China9	AL3194	D_Mal68	0.065	6.504	45,784	40,195	877630
D_China9	C_24T	D_AHusky91	0.0776	7.936	59,705	51,110	1,078553
D_China9	C_24T	D_Green	0.0926	9.696	61,461	51,046	1,133912
D_China9	C_24T	D_Husky	0.0799	8.709	65,471	55,786	1,150989
D_China9	C_24T	D_Husky89	0.1001	9.815	54,894	44,904	974829
D_China9	C_24T	D_Mal68	0.0743	7.814	59,571	51,330	1,113510
D_Port61	AL3194	D_AHusky91	0.0981	9.344	49,646	40,780	897394
D_Port61	AL3194	D_Green	0.0733	7.059	49,098	42,391	939,704
D_Port61	AL3194	D_Husky	0.0994	10.286	54,206	44,402	956,025
D_Port61	AL3194	D_Husky89	0.1006	9.791	43,556	35,593	784,550
D_Port61	AL3194	D_Mal68	0.1044	9.88	49,860	40,438	926,900
D_Port61	C_24T	D_AHusky91	0.1115	10.706	64,216	51,337	1,137,495
D_Port61	C_24T	D_Green	0.088	8.191	63,226	53,005	1,191,621
D_Port61	C_24T	D_Husky	0.1125	11.651	70,044	55,878	1,213,643
D_Port61	C_24T	D_Husky89	0.1181	11.113	57,125	45,064	1,005,524
D_Port61	C_24T	D_Mal68	0.1167	11.029	64,747	51,223	1,174,636
D_Portt71	AL3194	D_AHusky91	0.0949	9.146	48,273	39,904	877,094
D_Portt71	AL3194	D_Green	0.0814	8.348	48,157	40,909	918,351
D_Portt71	AL3194	D_Husky	0.0967	10.277	52,649	43,367	934,632
D_Portt71	AL3194	D_Husky89	0.1035	10.427	42,550	34,572	765,611
D_Portt71	AL3194	D_Mal68	0.1047	10.09	48,497	39,308	906,275
D_Portt71	C_24T	D_AHusky91	0.1122	10.895	62,687	50,038	1,112,724
D_Portt71	C_24T	D_Green	0.0994	9.888	62,205	50,964	1,165,580
D_Portt71	C_24T	D_Husky	0.1139	12.371	68,354	54,374	1,187,444
D_Portt71	C_24T	D_Husky89	0.124	11.829	55,911	43,573	982,083
D_Portt71	C_24T	D_Mal68	0.1214	11.729	63,196	49,518	1,149,527

Table 3.12 *D*(Outgroup, Asian or European Dogs, Arctic dogs, Arctic dogs). The Andean fox was used as an outgroup for these analyses. *n* corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks.

Pop 2	Pop 3	Pop 4	*D*	*Z*	ABBA	BABA	n
D_Portt71	D_AHusky91	D_Husky89	0.0000	0.001	37,394	37,394	954,618
D_Portt71	D_Mal68	D_Husky89	0.0005	0.040	47,184	47,141	988,232
D_Portt71	D_Husky	D_Mal68	0.0009	0.086	58,360	58,259	1,190,049
D_Port61	D_AHusky91	D_Mal68	0.0016	0.149	55,075	54,905	1,143,740
D_Port61	D_Mal68	D_Husky	0.0018	0.186	59,906	59,692	1,217,824
D_Portt71	D_Husky89	D_Husky	0.0026	0.250	42,139	41,923	1,014,889
D_Port61	D_Husky89	D_Mal68	0.0029	0.271	48,766	48,481	1,013,144
D_Portt71	D_AHusky91	D_Mal68	0.0046	0.412	53,778	53,282	1,117,362
D_Port61	D_AHusky91	D_Husky	0.0060	0.586	48,323	47,748	1,179,281
D_Portt71	D_AHusky91	D_Husky	0.0060	0.596	46,884	46,321	1,151,944
D_Port61	D_Husky89	D_AHusky91	0.0077	0.675	38,794	38,202	979,116
D_Port61	D_Husky89	D_Husky	0.0098	0.948	43,569	42,725	1,040,592
D_Portt71	D_Green	D_AHusky91	0.0197	1.780	53,880	51,794	1,133,450
D_Portt71	D_Green	D_Husky	0.0242	2.401	59,660	56,844	1,207,389
D_Portt71	D_Green	D_Mal68	0.0281	2.433	49,870	47,151	1,171,857
D_Portt71	D_Green	D_Husky89	0.0267	2.466	47,462	44,995	1,007,512
D_Port61	D_Green	D_Husky89	0.0283	2.625	48,938	46,241	1,033,241
D_Port61	D_Green	D_AHusky91	0.0301	2.628	55,922	52,657	1,160,778
D_Port61	**D_Green**	**D_Husky**	**0.0341**	**3.348**	**61,488**	**57,440**	**1,236,161**
D_Port61	**D_Green**	**D_Mal68**	**0.0375**	**3.355**	**51,677**	**47,947**	**1,199,625**
D_China8	D_Husky89	D_AHusky91	0.0017	0.158	37,346	37,218	959,372
D_China9	D_Mal68	D_AHusky91	0.0040	0.424	52,229	51,810	1,085,883
D_China9	D_Green	D_Husky89	0.0063	0.586	46,453	45,871	1,003,859
D_China8	D_Green	D_Mal68	0.0066	0.628	48,381	47,747	1,175,269
D_China8	D_Husky	D_AHusky91	0.0068	0.744	46,638	46,006	1,155,853
D_China8	D_Mal68	D_Husky	0.0071	0.780	58,261	57,441	1,193,177
D_China8	D_Husky	D_Husky89	0.0078	0.820	42,094	41,441	1,019,519
D_China9	D_Husky	D_Green	0.0081	0.846	57,408	56,484	1,178,134
D_China9	D_AHusky91	D_Husky	0.0085	0.957	45,878	45,108	1,120,029
D_China9	D_Mal68	D_Husky	0.0096	1.121	57,254	56,163	1,156,324
D_China8	D_Green	D_Husky	0.0124	1.296	58,444	57,014	1,211,379
D_China9	D_Husky	D_Husky89	0.0141	1.428	42,263	41,090	1,009,765
D_China9	D_AHusky91	D_Green	0.0152	1.493	52,310	50,746	1,106,207
D_China8	D_Mal68	D_AHusky91	0.0146	1.553	53,877	52,324	1,120,670
D_China8	D_Mal68	D_Husky89	0.0160	1.601	47,812	46,307	992,341
D_China8	D_Green	D_AHusky91	0.0193	1.872	53,594	51,565	1,137,595
D_China9	D_Husky89	D_AHusky91	0.0208	1.901	37,871	36,329	950,029
D_China9	D_Mal68	D_Green	0.0205	2.167	48,324	46,381	1,142,965
D_China9	D_Mal68	D_Husky89	0.0224	2.288	47,913	45,810	982,870
D_China8	D_Green	D_Husky89	0.0250	2.488	47,327	45,021	1,012,257

Table 3.13 Results of f_4 ratio analysis depicted in Fig. 3.2 with Z>2 (Table 3.5) for population codes).

PopA	PopO	PopX	PopC		PopA	PopO	PopB	PopC	Alpha	Std.Err	Z
CTVT	OUT	VDUA	VDPNGPM	:	CTVT	OUT	PCD	VDPNGPM	0.204	0.013	15.412
CTVT	OUT	VDUA	VDPNGEH	:	CTVT	OUT	PCD	VDPNGEH	0.194	0.013	14.816
CTVT	OUT	VDUA	BEA	:	CTVT	OUT	PCD	BEA	0.204	0.016	13.09
CTVT	OUT	VDUA	CSP	:	CTVT	OUT	PCD	CSP	0.180	0.016	11.361
CTVT	OUT	VDUA	VDIJ	:	CTVT	OUT	PCD	VDIJ	0.202	0.018	11.011
CTVT	OUT	VDUA	DEU	:	CTVT	OUT	PCD	DEU	0.197	0.018	10.884
CTVT	OUT	VDUA	DTM	:	CTVT	OUT	PCD	DTM	0.200	0.019	10.731
CTVT	OUT	VDUA	VDVLS	:	CTVT	OUT	PCD	VDVLS	0.177	0.017	10.671
CTVT	OUT	VDUA	BOX	:	CTVT	OUT	PCD	BOX	0.205	0.020	10.281
CTVT	OUT	VDUA	DCH	:	CTVT	OUT	PCD	DCH	0.181	0.018	9.878
CTVT	OUT	VDUA	CC	:	CTVT	OUT	PCD	CC	0.167	0.017	9.692
CTVT	OUT	VDUA	VDVLC	:	CTVT	OUT	PCD	VDVLC	0.159	0.016	9.692
CTVT	OUT	VDUA	VDIB	:	CTVT	OUT	PCD	VDIB	0.159	0.016	9.676
CTVT	OUT	VDUA	VDVHG	:	CTVT	OUT	PCD	VDVHG	0.154	0.017	9.148
CTVT	OUT	VDUA	DLB	:	CTVT	OUT	PCD	DLB	0.182	0.020	8.898
CTVT	OUT	VDUA	DGS	:	CTVT	OUT	PCD	DGS	0.198	0.023	8.702
CTVT	OUT	VDUA	DTI	:	CTVT	OUT	PCD	DTI	0.166	0.020	8.373
CTVT	OUT	VDUA	VDVCB	:	CTVT	OUT	PCD	VDVCB	0.138	0.017	8.041
CTVT	OUT	VDUA	NGSD	:	CTVT	OUT	PCD	NGSD	0.183	0.023	7.815
CTVT	OUT	VDUA	DAE	:	CTVT	OUT	PCD	DAE	0.128	0.020	6.375
CTVT	OUT	VDUA	DVN	:	CTVT	OUT	PCD	DVN	0.117	0.025	4.768

most PCD admixture. This is not surprising as these are closely related to Arctic breeds (Brown et al., 2015; Shannon et al., 2015). The f_4 ratio conducted above was not completely appropriate for this population (as it assumes close relatedness to Eurasian dogs; Fig. 3.2). We tested whether Alaskan village dogs had any PCD ancestry (interbred with pre-contact dogs) by computing every possible combination of the same f_4 ratio (Fig. 3.2) but using only Arctic breeds (now including Alaskan village dogs). We found that both Alaskan malamute and Greenland sledge dog have a significant amount of ancestry from PCD (Tables 3.14 and 3.15). might be the result of substructure among Arctic dogs. ADMIXTURE results supported the f_4 ratio analysis showing that the majority of modern American dog populations, including 138 village dogs from South America and multiple "native" breeds (e.g. Mexican hairless dogs and Louisiana catahoula leopard dogs), possess <4% PCD ancestry (Fig. 3.12).

We found varying degrees of PCD/Arctic ancestry in three individual Carolina dogs (0-33%; Fig. 3.12). Such signal might not have been detected by f_4 analysis as a result of the variable amount of ancestry in this population. This analysis also revealed an affinity between Chinook and the PCD/CTVT/Arctic population (12-15%;

Table 3.14 SNP array results of f_4 ratio analysis used to estimate proportion of PCD ancestry in Arctic breeds. Statistics with Z>2 are shown in bold.

PopX	PopC		PopB	PopC	alpha	std.err	z
AM	**VDUA**	:	**PCD**	**VDUA**	**0.122**	**0.021**	**5.882**
GSD	**VDUA**	:	**PCD**	**VDUA**	**0.132**	**0.024**	**5.503**
AM	**SH**	:	**PCD**	**SH**	**0.097**	**0.023**	**4.266**
GSD	**SH**	:	**PCD**	**SH**	**0.105**	**0.026**	**4.004**
GSD	**DAL**	:	**PCD**	**DAL**	**0.096**	**0.039**	**2.443**
AM	**DAL**	:	**PCD**	**DAL**	**0.076**	**0.036**	**2.095**
SH	VDUA	:	PCD	VDUA	0.028	0.02	1.413
DAL	VDUA	:	PCD	VDUA	0.043	0.034	1.268
DAL	SH	:	PCD	SH	0.02	0.03	0.651
GSD	AM	:	PCD	AM	0.008	0.026	0.311
AM	GSD	:	PCD	GSD	-0.007	0.027	-0.279
SH	DAL	:	PCD	DAL	-0.019	0.032	-0.601
VDUA	DAL	:	PCD	DAL	-0.043	0.037	-1.171
VDUA	SH	:	PCD	SH	-0.028	0.021	-1.351
DAL	AM	:	PCD	AM	-0.080	0.042	-1.893
DAL	GSD	:	PCD	GSD	-0.103	0.048	-2.165
SH	GSD	:	PCD	GSD	-0.116	0.033	-3.552
SH	AM	:	PCD	AM	-0.106	0.028	-3.825
VDUA	GSD	:	PCD	GSD	-0.152	0.032	-4.745
VDUA	AM	:	PCD	AM	-0.138	0.027	-5.138

Table 3.15 Genome-wide array results of f_4 ratio analysis used to estimate proportion of PCD ancestry in Arctic breeds. Statistics with Z>2 are shown in bold.

PopX	PopC		PopB	PopC	alpha	std.err	z
DMA	**DHU**	:	**PCD**	**DHU**	**0.114**	**0.027**	**4.209**
GSD	**DHU**	:	**PCD**	**DHU**	**0.122**	**0.028**	**4.41**
GSD	**DAL**	:	**PCD**	**DAL**	**0.116**	**0.031**	**3.743**
AM	**DAL**	:	**PCD**	**DAL**	**0.104**	**0.03**	**3.496**

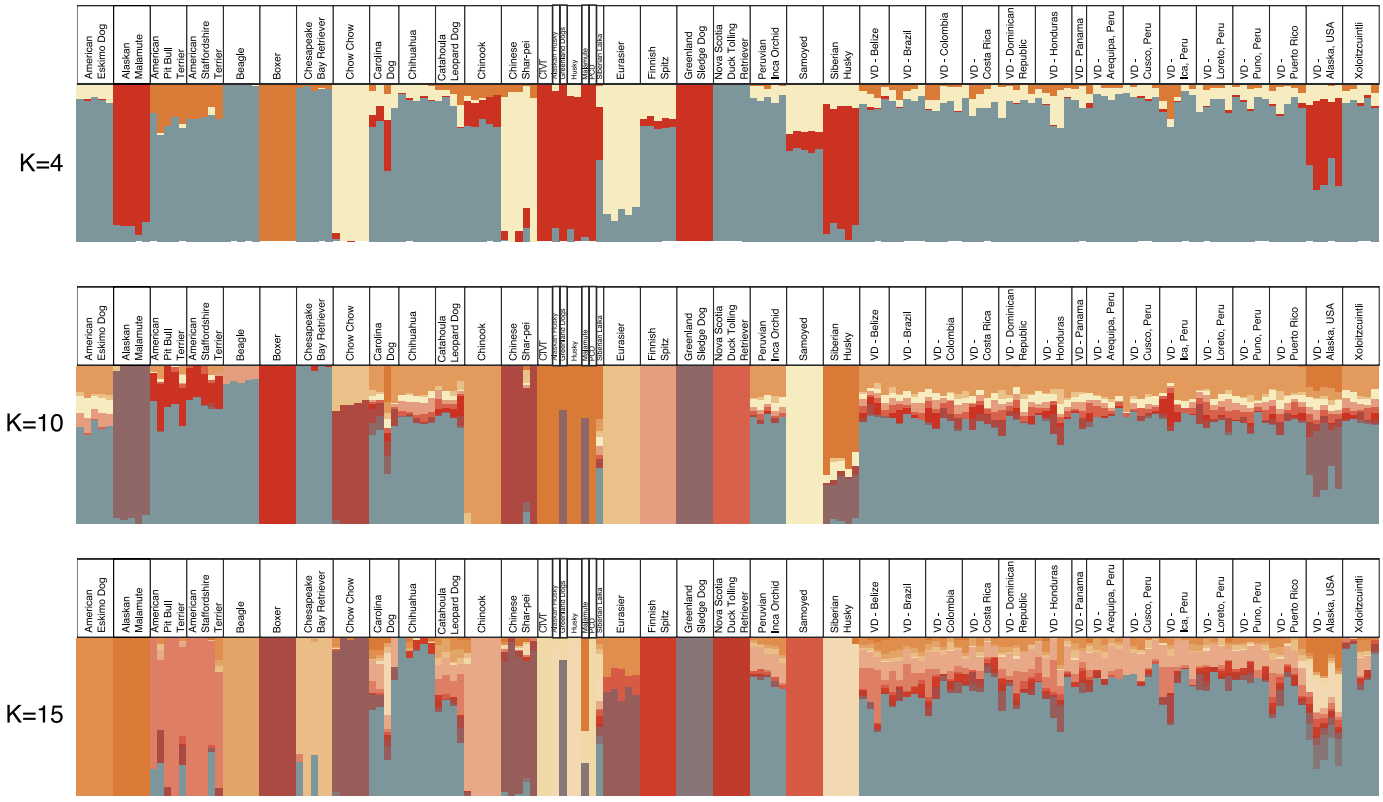


Fig. 3.12. ADMIXTURE results based on SNP array data for K = 4, 10 and 15. VD, Village Dog. Figure adapted from Ní Leathlobhair et al. (2018).

Fig. 3.12). This is not all that surprising given that Chinook dogs belong to the sledge dog grouping. With K = 4, however, it was not possible to distinguish between PCD and Arctic dog ancestry. The PCD/Arctic component in Carolina dogs and Chinook might have been introduced as a result of admixture with Arctic dogs and not from PCD. To test this we tried to separate PCD/Arctic ancestry with higher K values. Both K = 10 and K = 15, however, failed to differentiate PCD and Arctic dog ancestry but instead differentiated American Arctic dogs (Alaskan malamute and Greenland sledge dogs) from Eurasian Arctic dogs (Fig. 3.12).

3.11.7 Pre-contact ancestry in Eurasian dog populations

Returning to the mtDNA analysis, out of the 667 modern domestic dog mitochondrial genomes analysed in this chapter, only five modern samples were found carrying a pre-contact mtDNA haplogroup: a Terrier cross from San Juan del Sur, Nicaragua (GenBank accession: KU291094), a Chihuahua (GenBank accession: EU408262), a Japanese Spitz (GenBank accession: EU789755), a non-breed dog from Shanxixian, China (GenBank accession: EU789669) and a non-breed dog from Laem Ngop, Thailand (GenBank accession: EU789664). Two out of five of these sequences derive from dogs in the Americas. Three, however, are from East Asia. This suggests a very low frequency (~2.5%) of the PCD mtDNA haplogroup in East Asia .

All five of the aforementioned modern samples cluster together with ancient Mexican dogs from Mayapan. We inferred the age of the MRCA between the Mayapan dogs and the five modern dogs identified as monophyletic with PCD. We found that these five modern dogs diverged from the Mayapan dogs between 6,289-9,865 years ago (95% HPDI) suggesting that their divergence postdates the flooding of the land bridge between Western and Eastern Beringia. Given this result, it is less likely that the mtDNA haplotype of these dogs originated in Eurasia. These results instead suggest that dogs carrying PCD ancestry were transported from the Americas into East Asia. This most likely took place during recent times and could be linked to the creation of hairless dog breeds in Asia (Drögemüller et al., 2008). However, further work is required to test these possibilities.

Table 3.16 *D*-statistics for *D*(Outgroup, North American wolf or Coyote or Taimyr wolf, PCD, PCD). *n* corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks

Pop2	Pop3	Pop4	*D*	*Z*	BABA	ABBA	n
C_Cal	AL3194	AL3223	0.0152	1.460	5,681	5,511	285,638
C_Cal	AL3223	AL3194	-0.0152	-1.460	5,511	5,681	285,638
C_MidW	AL3194	AL3223	0.0035	0.277	3,443	3,419	181,671
C_MidW	AL3223	AL3194	-0.0035	-0.277	3,419	3,443	181,671
W_Mex1	AL3194	AL3223	-0.0130	-1.086	7,394	7,588	282,977
W_Mex1	AL3223	AL3194	0.0130	1.086	7,588	7,394	282,977
W_Yellow1	AL3194	AL3223	-0.0083	-0.785	7,501	7,626	281,684
W_Yellow1	AL3223	AL3194	0.0083	0.785	7,626	7,501	281,684
W_Yellow2	AL3194	AL3223	-0.0067	-0.622	7,513	7,615	282,189
W_Yellow2	AL3223	AL3194	0.0067	0.622	7,615	7,513	282,189
TAI	AL3194	AL3223	-0.0001	-0.005	3,026	3,026	116,212
TAI	AL3223	AL3194	0.0001	0.005	3,026	3,026	116,212

3.11.8 Introgression from wild American canids

The relatedness between the CTVT founder dog and PCD, as well as the time of CTVT origin, is compatible with the possibility that the CTVT founder dog was in North America. To further assess the plausibility of this scenario, we quantified introgression from North American endemic canids (coyotes and North American wolves) into pre-contact dogs, modern Arctic dogs, and the CTVT founder dog.

The CTVT founder dog showed weak evidence of admixture with coyotes, but does not appear to have admixed with North American wolves (Fig. 3.13).
These analyses also indicated that PCDs from Port au Choix and Weyanoke (~2× and ~0.5× coverage, respectively; Fig. 3.1a) share a significant number of derived alleles with coyotes and North American wolves, indicative of admixture (Fig. 3.13 and 3.14). We also tested for extra admixture from wild canids into our higher coverage PCD genomes (AL3194; AL3223) but found no evidence of this (Table 3.16).

PCA suggests that the Koster dog (AL2135; ~9,900 cal BP) was admixed with wild canids. We found borderline significant results ($|Z|>2$; Table 3.17) suggestive of admixture from Coyote into AL2135. This sample, however, was very low coverage (only ~17K SNPs were called). Its placement on the PCA and this positive admixture

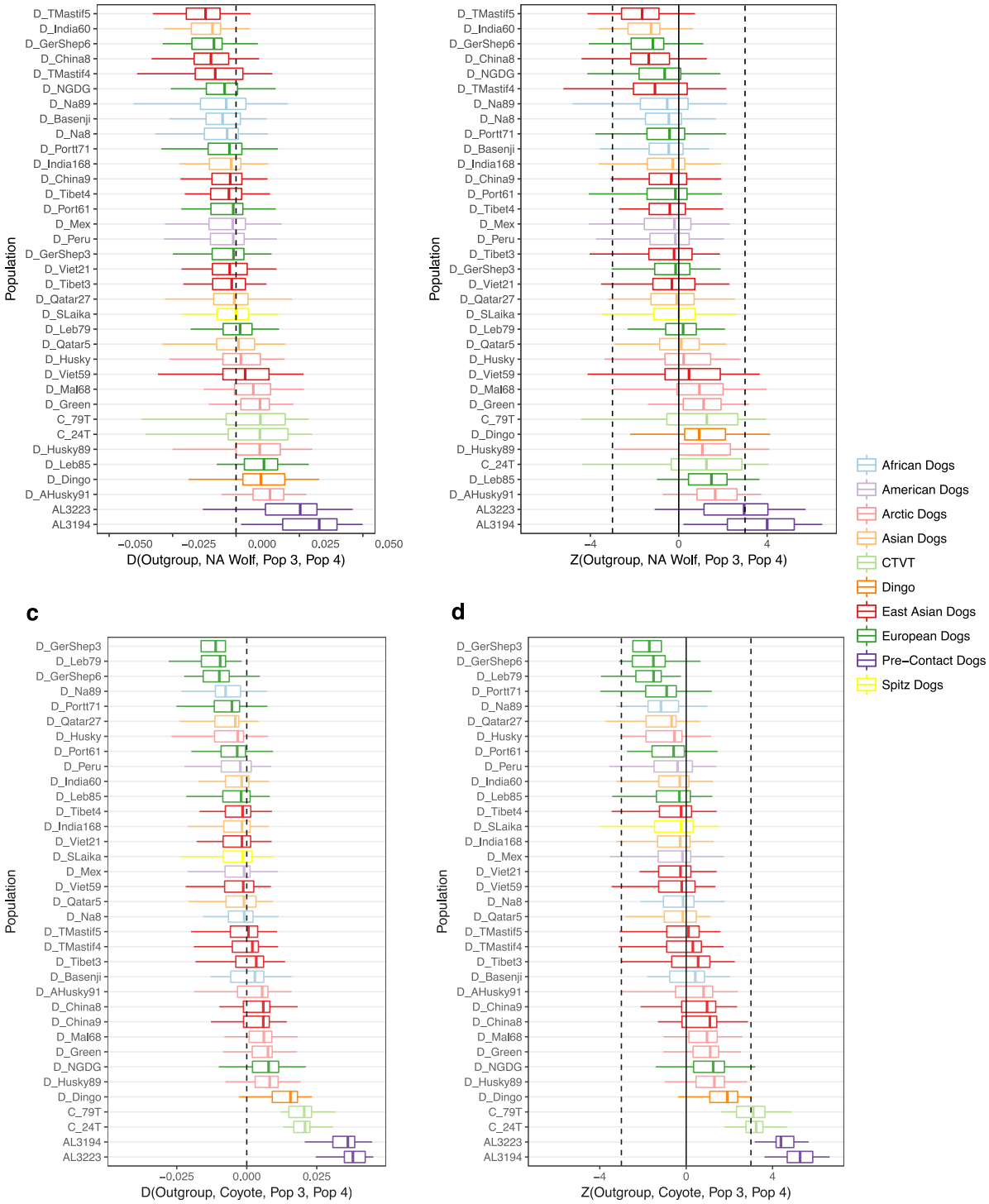


Fig. 3.13. Box plots representing *D*-statistics for every combination of **a.** *D*(Outgroup, North American Wolf, Pop3, Pop4) **c.** *D*(Outgroup, Coyote, Pop3, Pop4) where Population 3 (Pop 3) and Population 4 (Pop 4) represent every possible combination of dog populations. Box plots representing significance of *D*-statistics are shown in **b.** and **d.** The y-axis represents Pop 3. Positive values support a close relationship between Pop 3 and the test population (Coyote or North American Wolf) while negative support a close relationship between Pop 4 and the test population. Figure adapted from Ní Leathlobhair et al. (2018).

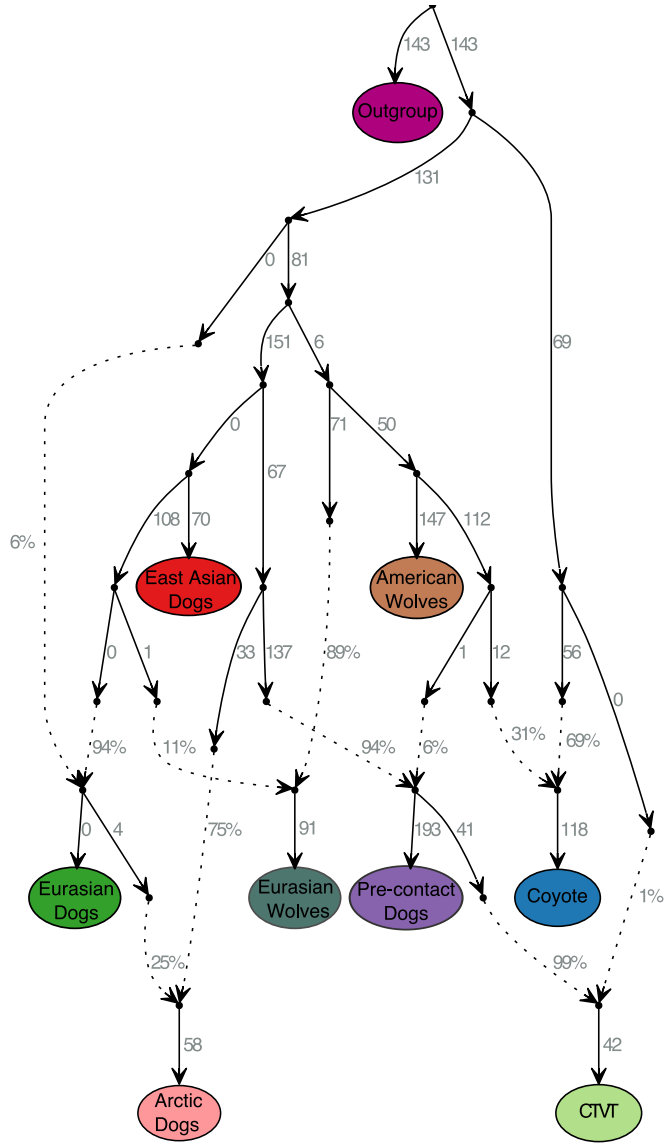


Fig. 3.14. Qpgraph model with admixture fractions including West Eurasian dogs (Portuguese village dogs), East Asian dogs (Vietnamese village dogs), pre-contact dogs (PCD) (Port au Choix, AL3194 and Weyanoke Old Town, AL3223), Arctic dogs (Malamute), CTVT (79T and 24T), Eurasian wolves from Spain and Portugal, North American wolves from Yellowstone, and coyotes as an outgroup. Figure adapted from Ní Leathlobhair et al. (2018)

Table 3.17 *D*-statistics for *D*(Outgroup, North American wolf or Coyote or Taimyr wolf, PCD, PCD). The Andean fox was used as an outgroup for these analyses. *n* corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks. Statistics with *Z*>3 and *Z*<-3 are shown in bold.

Pop2	Pop3	Pop4	*D*	*Z*	BABA	ABBA	n
W_Mex1	AL2135	AL3194	0.0690	1.024	112	98	2,762
W_Mex1	AL3194	AL2135	-0.0690	-1.024	98	112	2,762
W_Yellow1	AL2135	AL3194	0.0808	1.304	124	106	2,758
W_Yellow1	AL3194	AL2135	0.0808	-1.304	106	124	2,758
W_Yellow2	AL2135	AL3194	0.0792	1.248	120	102	2,764
W_Yellow2	AL3194	AL2135	-0.0792	-1.248	102	120	2,764
C_Cal	AL2135	AL3194	-0.0004	-0.006	87	87	2,794
C_Cal	AL3194	AL2135	0.0004	0.006	87	87	2,794
C_MidW	**AL2135**	**AL3194**	**-0.1814**	**-2.034**	**43**	**62**	**1,863**
C_MidW	**AL3194**	**AL2135**	**0.1814**	**2.034**	**62**	**43**	**1,863**
TAI	AL2135	AL3194	0.1284	1.217	53	41	1,098
TAI	AL3194	AL2135	-0.1284	-1.217	41	53	1,098

signal might therefore a result to this low coverage.

It had been suggested that some North American wolves obtained a mutation leading to black coat colour possibly via admixture with early American dogs (Anderson et al., 2009). However, no evidence was found for the CBD103$^{\triangle G23}$/ K^{B} mutation in either of the two higher-coverage ancient PCDs analysed (AL3194 and AL3223; Table 3.3) or in CTVT. Additional ancient genomes are necessary to determine if this allele was present in the PCD population.

3.11.9 Introgression from the archaic Taimyr wolf

Skoglund et al. (2015) previously presented evidence for introgression from an ancient Siberian wolf lineage into modern dog breeds from northeast Siberia and Greenland. Testing for Taimyr admixture in our ancient and modern dog sample set using *D*-statistics (Section 3.9.7), we found few values with $|Z|$>3 (AL3194 and malamute; Table 3.18). Lowering the threshold to $|Z|$>2.5, we found admixture from the Taimyr wolf into PCD (both AL3194 and AL3223) as well as in all Arctic dogs (husky, Greenland sledge dog, and malamute) and CTVT. No evidence was found for excess admixture from the Taimyr wolf into either CTVT, PCD or Arctic populations (Table 3.19). This suggests that the Taimyr wolf introgression into Arctic

Table 3.18 *D*-statistics for *D*(Outgroup, Taimyr, PCD/Arctic dogs, European, Asian, or Arctic dogs). *n* corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks. Statistics with *Z*>3 and *Z*<-3 are shown in bold.

Pop2	Pop3	Pop4	*D*	*Z*	BABA	ABBA	n
TAI	**AL3194**	**D_India168**	**-0.0274**	**-3.377**	**13,074**	**13,812**	**297,919**
TAI	**AL3194**	**D_Peru**	**-0.0267**	**-3.293**	**17,139**	**18,079**	**384,525**
TAI	**AL3194**	**D_India60**	**-0.0247**	**-3.254**	**17,383**	**18,264**	**372,784**
TAI	**AL3194**	**D_Na8**	**-0.0256**	**-3.199**	**16,614**	**17,489**	**366,726**
TAI	**D_Mal68**	**D_Na8**	**-0.0206**	**-3.127**	**19,530**	**20,350**	**459,311**
TAI	**AL3194**	**D_Port61**	**-0.0236**	**-3.006**	**17,398**	**18,239**	**387,777**
TAI	D_Mal68	D_Peru	-0.0195	-2.856	20,182	20,983	482,639
TAI	D_Mal68	D_India60	-0.0189	-2.789	20,930	21,735	465,930
TAI	D_Mal68	D_Port61	-0.0204	-2.768	20,358	21,206	485,553
TAI	AL3194	D_Viet59	-0.0227	-2.753	13,212	13,824	308,811
TAI	D_Green	D_Na8	-0.0202	-2.674	20,383	21,224	465,676
TAI	D_Husky89	D_India168	-0.0205	-2.664	13,042	13,588	314,547
TAI	D_Green	D_India60	-0.0198	-2.607	21,505	22,372	472,462
TAI	D_Husky89	D_Port61	-0.0215	-2.571	17,284	18,043	413,104
TAI	D_Husky89	D_Peru	-0.0201	-2.525	17,191	17,896	410,031
TAI	D_Green	D_Port61	-0.0197	-2.506	21,204	22,058	492,659
TAI	AL3194	D_Husky	-0.0199	-2.498	16,444	17,111	394,720
TAI	AL3194	D_SLaika	-0.0194	-2.485	16,821	17,487	379,149
TAI	D_Green	D_India168	-0.0178	-2.481	15,728	16,296	377,110
TAI	D_Mal68	D_Na89	-0.0195	-2.481	12,195	12,680	289,545
TAI	D_Green	D_Peru	-0.0194	-2.455	20,782	21,605	488,730
TAI	D_Husky89	D_India60	-0.0186	-2.44	17,456	18,117	394,520
TAI	AL3194	D_Mex	-0.0193	-2.426	17,007	17,675	380,509
TAI	AL3194	D_Viet21	-0.0204	-2.419	13,088	13,634	303,723
TAI	AL3194	D_Tibet3	-0.0184	-2.41	16,776	17,403	375,015
TAI	AL3194	D_Portt71	-0.0185	-2.366	16,997	17,637	379,175

dogs suggested in Skoglund et al. (2015) may have taken place after the PCD, Arctic dog and CTVT lineage diverged from Eurasian dogs but before the divergence of the Arctic dog and PCD/CTVT lineages.

3.12 Discussion

In this section, I reconstruct a narrative of CTVT emergence and pre-contact dog population history consistent with results outlined above. I discuss the limitations of this work results and identify further research questions arising from the results of this chapter.

3.12.1 Emergence of CTVT

This analysis infers an upper bound for the emergence of CTVT, 8,225 years ago based on mutation count data from a case of direct transmission of CTVT from mother to pup during birth. Although we have determined that CTVT first arose in a dog that was closely related to PCDs, this analysis does not pinpoint the precise location in which CTVT originated. As coyotes are currently restricted to North America, evidence of coyote introgression (Fig. 3.13c) suggests the CTVT founder dog may have been in North America; however, due to sampling limitations, we were unable to ascertain the degree of coyote ancestry in ancient PCD-related dogs outside of America (such as the Zhokhov Island dogs in Northern Siberia). Genome-wide data from the ancient Zhokhov dog population will provide valuable insight into the geographic origin of CTVT as well as PCD population history.

A reconstructed phylogeography of the CTVT clone, described in Chapter 2 (Section 2.6.5), rather indicates that the earliest dispersals of the lineage were in Asia, suggesting that CTVT originated in Northern or Central Asia. Further, the phylogeny suggests a later, post-contact introduction of the modern CTVT lineage found in the Americas followed by rapid dispersal (Fig. 2.7; Fig. 2.15). However, we cannot exclude the possibility that the clone first arose in America and then dispersed early into Asia before being reintroduced to America.

Both scenarios raise the intriguing possibility that CTVT may have contributed to PCD demise. Intriguingly, a recent study proposed that infectious disease may have constrained the spread of dogs into South America and explains the relatively

Table 3.19 *D*-statistics for *D*(Outgroup, Taimyr, PCD or Arctic dogs, CTVT or Arctic dogs or PCD). *n* corresponds to the number of SNPs where all populations have data. Statistics indicate that there has been no extra admixture from the Taimyr wolf into other populations.

Pop2	Pop3	Pop4	*D*	*Z*	BABA	ABBA	n
TAI	AL3194	C_24T	-0.0239	-2.258	10,176	10,674	380,938
TAI	AL3194	C_79T	-0.0267	-2.364	8,948	9,438	346,188
TAI	AL3194	D_AHusky91	-0.0126	-1.58	15,165	15,554	370,227
TAI	AL3194	D_Green	-0.0059	-0.681	15,468	15,650	387,439
TAI	AL3194	D_Husky	-0.0199	-2.498	16,444	17,111	394,720
TAI	AL3194	D_Husky89	-0.0076	-0.915	13,258	13,462	321,758
TAI	AL3194	D_Mal68	-0.0025	-0.312	15,225	15,301	382,193
TAI	C_24T	AL3194	0.0239	2.258	10,674	10,176	380,938
TAI	C_24T	C_79T	-0.0014	-0.654	5,721	5,737	441,374
TAI	C_24T	D_AHusky91	0.0045	0.587	19,419	19,247	461,923
TAI	C_24T	D_Green	0.009	1.179	19,495	19,147	483,404
TAI	C_24T	D_Husky	-0.0028	-0.381	20,900	21,018	493,025
TAI	C_24T	D_Husky89	0.0071	0.929	17,107	16,865	405,779
TAI	C_24T	D_Mal68	0.0102	1.321	19,467	19,075	476,690
TAI	C_79T	AL3194	0.0267	2.364	9,438	8,948	346,188
TAI	C_79T	C_24T	0.0014	0.654	5,737	5,721	441,374
TAI	C_79T	D_AHusky91	0.0041	0.539	17,112	16,972	416,326
TAI	C_79T	D_Green	0.0062	0.786	17,041	16,832	435,794
TAI	C_79T	D_Husky	-0.0021	-0.279	18,489	18,566	446,651
TAI	C_79T	D_Husky89	0.0081	1.039	14,690	14,454	358,654
TAI	C_79T	D_Mal68	0.0113	1.419	17,079	16,698	429,798
TAI	D_AHusky91	AL3194	0.0126	1.58	15,554	15,165	370,227
TAI	D_AHusky91	C_24T	-0.0045	-0.587	19,247	19,419	461,923
TAI	D_AHusky91	C_79T	-0.0041	-0.539	16,972	17,112	416,326
TAI	D_AHusky91	D_Green	0.0083	1.034	17,900	17,607	470,340
TAI	D_AHusky91	D_Husky	-0.0081	-1.105	15,523	15,775	478,592
TAI	D_AHusky91	D_Husky89	0.0009	0.115	12,481	12,458	394,759
TAI	D_AHusky91	D_Mal68	0.0072	1.009	18,163	17,904	463,584
TAI	D_Green	AL3194	0.0059	0.681	15,650	15,468	387,439
TAI	D_Green	C_24T	-0.009	-1.179	19,147	19,495	483,404
TAI	D_Green	C_79T	-0.0062	-0.786	16,832	17,041	435,794
TAI	D_Green	D_AHusky91	-0.0083	-1.034	17,607	17,900	470,340
TAI	D_Green	D_Husky	-0.0137	-1.839	19,216	19,750	501,352
TAI	D_Green	D_Husky89	-0.0025	-0.296	15,501	15,578	416,390
TAI	D_Green	D_Mal68	-0.0002	-0.029	16,258	16,266	485,928
TAI	D_Husky	AL3194	0.0199	2.498	17,111	16,444	394,720
TAI	D_Husky	C_24T	0.0028	0.381	21,018	20,900	493,025
TAI	D_Husky	C_79T	0.0021	0.279	18,566	18,489	446,651
TAI	D_Husky	D_AHusky91	0.0081	1.105	15,775	15,523	478,592
TAI	D_Husky	D_Green	0.0137	1.839	19,750	19,216	501,352
TAI	D_Husky	D_Husky89	0.0123	1.655	14,207	13,860	419,790
TAI	D_Husky	D_Mal68	0.013	1.913	19,767	19,258	494,030
TAI	D_Husky89	AL3194	0.0076	0.915	13,462	13,258	321,758
TAI	D_Husky89	C_24T	-0.0071	-0.929	16,865	17,107	405,779
TAI	D_Husky89	C_79T	-0.0081	-1.039	14,454	14,690	358,654
TAI	D_Husky89	D_AHusky91	-0.0009	-0.115	12,458	12,481	394,759
TAI	D_Husky89	D_Green	0.0025	0.296	15,578	15,501	416,390
TAI	D_Husky89	D_Husky	-0.0123	-1.655	13,860	14,207	419,790
TAI	D_Husky89	D_Mal68	0.0043	0.587	15,803	15,667	408,271
TAI	D_Mal68	AL3194	0.0025	0.312	15,301	15,225	382,193
TAI	D_Mal68	C_24T	-0.0102	-1.321	19,075	19,467	476,690
TAI	D_Mal68	C_79T	-0.0113	-1.419	16,698	17,079	429,798
TAI	D_Mal68	D_AHusky91	-0.0072	-1.009	17,904	18,163	463,584
TAI	D_Mal68	D_Green	0.0002	0.029	16,266	16,258	485,928
TAI	D_Mal68	D_Husky	-0.013	-1.913	19,258	19,767	494,030
TAI	D_Mal68	D_Husky89	-0.0043	-0.587	15,667	15,803	408,271

late appearance of dogs on the continent (~6,000 years after their arrival in North America) (Mitchell, 2017). Given that CTVT is a contagious allograft, it is possible that it more effectively colonised and caused more severe morbidity in hosts with which the tumour shared genetic affinity. Previous analyses of heterozygosity in the founder have suggested that CTVT first emerged in a relatively inbred population of dogs as an opportunistic pathogen (Murchison et al., 2014; Murgia et al., 2006; Rebbeck et al., 2009). It is possible that the success of CTVT as a 'coloniser' relied on a reservoir of inbred individuals to host the tumour in its early stages; early competition for hosts within a small population of dogs may even have facilitated pathogenic genome adaptation in CTVT.

3.12.1.1 Assumptions underlying time-of-origin estimate

The model I have used to derive a somatic mutation rate for CTVT relied on the following assumptions: (*i*) the mother's tumour (609T) was seeded either by a single CTVT cell or a single clone of identical CTVT cells (i.e. monoclonal seeding), (*ii*) the age estimate provided for the pup (Dog 608) is accurate, (*iii*) the mother (Dog 609) was infected with CTVT at the time of the heat cycle during which she conceived the pup (Dog 608), (*iv*) the contribution of variable mutation opportunity to our estimates was negligible and the N[C>T]G rate has remained constant, (*v*) the contribution of back-mutations to our estimates was negligible, (*vi*) estimates of the total somatic mutation burden in CTVT are accurate, (*vii*) mutation false discovery rate in tumour-unique and tumour-only variant sets was equivalent.

In addition, I based my estimate of the CTVT mutation rate on a single observation of transmission. Future studies could address this, and account for sampling variation, by measuring mutation rates either across other samples derived from direct transmissions or spatially separated samples of the same tumour. Multi-region sampling and sequencing of tumour biopsies derived from the mother and pup sampled in this study would be useful in confirming clonal and subclonal populations within the tumours and could stand to refine our estimate of CTVT origin.

Despite these limitations, this analysis provides a plausible estimate of the CTVT somatic mutation rate, and is comparable with clock-like mutation rates observed in human cancers (Alexandrov et al., 2015).

3.12.2 Origin of pre-contact dogs

The earliest evidence of dogs in the North American archaeological record, dates to ~10,000 years ago (Perri et al., 2018). Our molecular clock analysis indicates that the PCD lineage appeared ~6,500 years after North American human lineages (Raghavan et al., 2015). A time frame in which dogs were brought to the Americas several thousand years after the first human migration (Moreno-Mayar et al., 2018), before the flooding of the Bering land bridge ~11,000 years ago (Jakobsson et al., 2017), would be compatible with both the archaeological record and the PCD divergence time estimate.

Although this study does not address the question of domestication, the findings of this chapter are broadly consistent with a single domestication followed by an early split into three geographically isolated groups, in line with recent studies (Botigué et al., 2017; Loog et al., 2018) and in contrast to the recent suggestion that extant global dog populations arose from two independent domestications (Frantz et al., 2016).

An in-depth exploration of dog history in the Americas will require more complete genomes along with higher coverage genomes, both nuclear and mitochondrial, from ancient and modern dogs. Notably, our sampling did not include dogs from sites associated with the Thule culture; these samples would be especially interesting considering that modern American Arctic dogs are plausibly the descendants of dogs introduced by the Thule. The seven pre-contact nuclear genomes generated as part of this study were sequenced to (0.005 to 2×), far lower than the coverage achieved in other aDNA studies (Botigué et al., 2017; Frantz et al., 2016), making it difficult to comprehensively genotype variants and limiting the types of analyses that could be applied.

3.12.3 Evolutionary history of pre-contact dogs

The evolutionary history of dogs in North America is characterised by multiple introductions and replacements (Fig. 3.15a) This analysis suggests, as discussed, that an initial dog population entered North America and then dispersed throughout the Americas, where it remained isolated for at least 9,000 years. Within the past 1,000 years, however, at least three independent migrations of dogs into the Americas

have occurred (Fig. 3.15a). The first may have consisted of Arctic dogs that arrived with the Thule culture ~1,000 years ago (Brown et al., 2015). Then, beginning in the 15th century, Europeans brought a second wave of dogs that appear to have almost completely replaced native dogs. Lastly, Siberian huskies were introduced to the American Arctic during the Klondike or Alaskan gold rush (Derr, 2005). As a result of these more recent introductions, the present-day population of American dogs is largely derived from Eurasian breeds.

Based on our results of our analyses (Section 3.11.5), it is likely that the Eurasian Arctic dogs that were recently brought into the Americas, as far as Greenland, originated from a population that was more closely related to PCD dogs than other Arctic dogs. The high degree of mtDNA divergence detected within ancient (~9,000 BP) Eurasian Arctic dogs from Zhokhov Island suggests that ancient substructure with Arctic dogs is a plausible scenario.

3.12.4 Disappearance of pre-contact dogs

The most puzzling aspect of PCD population history is their almost total disappearance. The results presented in this chapter suggest that European dogs almost completely replaced native dog lineages in the Americas. The decline of pre-contact dogs may have been influenced by changing cultural preferences, possibly influenced by the arrival of Europeans, or by persecution of native dogs (Derr, 2005). Similar to indigenous human populations in the Americas (Lindo et al., 2016), native dogs may have experienced high mortality rates during the early contact period as a result of newly introduced infectious diseases to which pre-contact dogs were susceptible (Velasco-Villa et al., 2017). It is also worth noting that while our mitogenome evidence suggests that PCD-like dogs were in North-East Siberia ~9,000 years ago, dogs with a close affinity to the PCD clade have not been sampled in modern Siberia.

However, it is plausible that the modern American Arctic dogs included in our analysis, such as Alaskan malamutes and Greenland dogs, are the descendants of dogs introduced to the Americas during the Thule expansion. Alternatively, Thule dogs may form part of the pre-contact dog lineage and the modern American Arctic dogs sampled in this study may be the descendants of Eurasian Arctic dogs, many of which were recently introduced during the 19th-century Alaskan gold rush and as

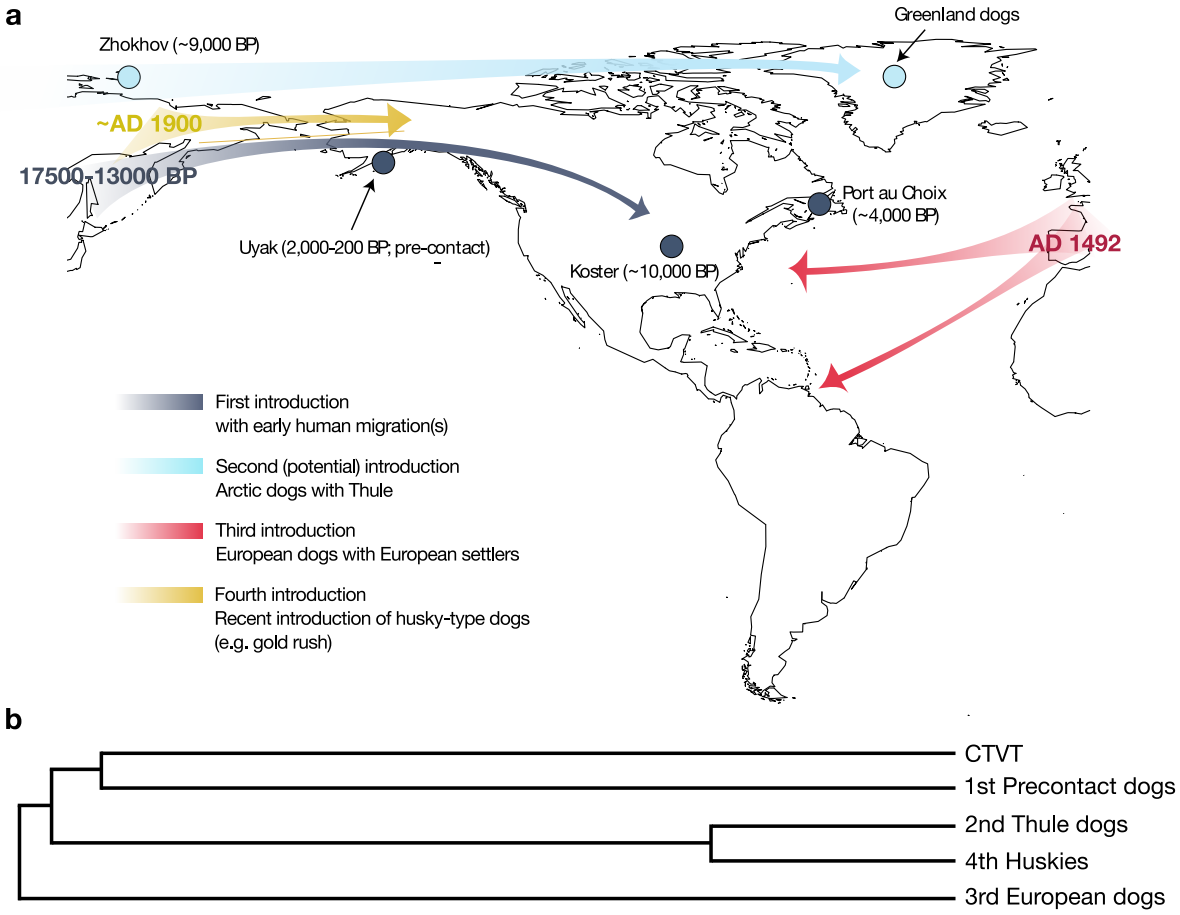


Fig. 3.15. a. A map depicting multiple migrations of dogs into the Americas consistent with archaeological and genetic data. Note that these arrows do not depict exact geographic migration routes. Key archaeological sites from which dog remains were analysed in this chapter are labelled. BP, before present. AD, Anno Domini. Figure adapted from Ní Leathlobhair et al. (2018). **b.** Tree schematic of four independent dog introductions to the Americas. Branch lengths are not meaningful.

sledge dog racing stock. In either case, the descendants of pre-contact or Thule dogs would both have admixed with Eurasian dogs. Finally, it is also possible that, similar to human populations (Ebenesersdóttir et al., 2018), extensive genetic drift resulting from a founder effect could explain the divergence of modern Arctic dogs from their pre-contact dog source population. This hypothesis has not yet been tested, however.

Mitochondrial ancestry analyses did detect the presence of the PCD mitochondrial haplotype in three East Asian dogs (Section 3.11.6). Multiple scenarios could explain the finding of East Asian dogs within the PCD clade: for instance, the clade to which these East Asian samples belong could have diverged from pre-contact dogs prior to their introduction into the Americas or there were back and forth migratory movements between the continents after the flooding of the land bridge between Western and Eastern Beringia ~11,000 years ago.

A major direction for future data collection would be to fill in sampling gaps in extant dog populations. We cannot rule out that some larger degree of PCD ancestry may remain in American dogs that have not yet been sampled. A further survey of the diversity and ancestry of dogs in North-East Asia, the North American Arctic and South America is warranted with special emphasis on 'indigenous' dog groups.

3.12.5 Summary

Pre-contact dogs represent a relatively unexplored evolutionary branch of the dog population. This lineage could be invaluable in understanding the history of dog domestication and genomic changes associated with early population bottlenecks.

Chapter 4

Investigating the transmission of urogenital carcinoma in California sea lions

4.1 Chapter abstract

This chapter deviates from the theme of the origin and evolution of the Canine Transmissible Venereal Tumour to consider the emergence of contagious cancer in other animal species, a possibility I discussed in the introduction to this thesis (Section 1.2.3). During the four years since the beginning of my PhD, six further transmissible cancers were discovered in multiple species (Metzger et al., 2015, 2016; Pye et al., 2016). These findings raise the possibility that other cancers could be due to transmission of cancer cells, specifically those that occur at high prevalence and those with a plausible route of transmission. Recent studies have also shown that transmissible cancer cells can transmit within marine environments (Metzger et al., 2015, 2016) and that these rogue cells might be common infectious agents.

Urogenital carcinoma is a highly metastatic cancer in California sea lions (*Zalophus californianus*) and is one of the most commonly observed cancers in wildlife. Although a putative viral agent (Otarine herpesvirus-1, OtHV-1) has been suggested and is associated with lesions, a causative role for this virus has not been confirmed. The genital localisation of primary tumours suggests the possibility that coital transmission of an infectious agent could underlie this disease.

Here, I investigate the etiology of a highly prevalent cancer in sea lions and, specifically, the possibility that UGC might be clonally transmitted by testing for genetic differences between tumour and host cells. Analysis of mitochondrial DNA control region sequences in seven matched tumour and host pairs confirmed that tumour

genotypes were identical to those of their matched hosts and did not show similarity with tumours from other individuals. Thus our findings suggest that urogenital carcinoma in California sea lions is not clonally transmitted, but rather that tumours derive from transformed host cells.

4.2 Publications associated with this chapter

This study was published in *Wellcome Open Research* as a Research Note on 27th June 2017 (Ní Leathlobhair et al. (2017); see Appendix 2 for the full reference).

4.3 Introduction

Cancer is widespread in animal populations and the reported incidence of neoplasia in wildlife species is increasing (Pesavento et al., 2018). Global climate change and anthropogenic activities environmental toxins such as plastics and mutagens. There are obvious challenges in diagnosing and monitoring cancer in wildlife species.

Urogenital carcinoma (UGC) is the most commonly observed neoplasm in California sea lions (Browning et al., 2015). This cancer was first reported in sea lions on the west coast of North America in 1979 (Gulland et al., 1996), and over a fourteen-year period, from 1998 to 2012, the disease was found in 26 per cent of adult animals examined post-mortem at The Marine Mammal Center, California (Browning et al., 2015). The prevalence of the disease in the wider sea lion population is unknown UGC affects both male and female animals, and is most frequently found in sexually mature adults (Colegrove et al., 2009; Lipscomb et al., 2010). The disease typically presents with extensive multi-organ metastases; however, primary lesions involving the genital epithelium can usually be identified (Lipscomb et al., 2000).

Three aetiological factors have been proposed for the development of UGC: infection, host genetics, and environmental factors. Otarine herpesvirus-1 infection, a gammaherpesvirus related to Kaposi's sarcoma-linked human herpesvirus-8 (King et al., 2002; Lipscomb et al., 2000) has been associated with UGC (Buckles et al., 2006; King et al., 2002; Lipscomb et al., 2000); however, this virus has not been confirmed as a causative agent. An association between UGC and genital bacterial infection has also

Fig. 4.1. Urogenital Carcinoma presentation in California Sea Lions. Primary masses in the urinary and genital tract can be observed as well as metastatic masses in the adrenals, ovaries, liver, kidney, spleen and serosal surfaces. Images are reproduced with permission from Frances Gulland (The Marine Mammal Center, Sausalito, CA).

been proposed (Johnson et al., 2006). Genetic studies have indicated that individuals with high parental relatedness (Acevedo-Whitehouse et al., 2003), homozygosity at the heparanase-2 (*HSPE2*) locus (Browning et al., 2014), or one or more copies of the *Zaca-DRB.A* MHC class II locus (Bowen et al., 2005) have increased risk of UGC. Environmental contaminants, such as organochlorines, have also been proposed as causative agents in UGC carcinogenesis (Ylitalo et al., 2005).

Several features of UGC are compatible with the possibility that this cancer is clonally transmissible: epidemiological observations of UGC are consistent with an infectious aetiology for the disease (Gulland et al., 1996); and, in particular, its genital localisation could provide a coital route of transmission (Buckles et al., 2007), as observed in CTVT.

In the introduction, I have covered features of UGC pathogenesis that make this cancer a reasonable suspect for clonal transmission. The straightforward aim of this chapter was to test whether urogenital carcinoma in sea lions could be a transmissible cancer in collaboration with researchers at The Marine Mammal Center.

4.4 Methods

4.4.1 Sample collection

This study was approved by The Marine Mammal Center Institutional Animal Care and Use Committee (Sausalito, CA) and the National Marine Fisheries Service MMPA (permit number 18786).

Tissues from seven wild stranded adult California sea lions were collected at The Marine Mammal Center, Sausalito, CA. Complete gross and histopathological examinations were performed on each animal to confirm UGC diagnosis. Tumour (metastasis) and host tissue (liver or muscle) biopsies were collected into RNAlater during post-mortem examination and were stored at -70°C until processing. Jinhong Wang (Cambridge) and Barbie Halaska assisted with sample collection.

4.4.2 DNA extraction

Representative tissue sampled from tumour and host biopsies was used for DNA extraction using the Qiagen DNeasy Blood and Tissue Kit according to manufacturer's instructions. DNA was quantified using a Qubit 2.0 fluorometer.

4.4.3 PCR

A 1289 bp fragment of the mtDNA control region was amplified using the following primer set, described by Wolf et al. (2007): Dloop_F, 5′ CCTCCCTAAGACTCAAGGAAGAA 3′; Dloop_R, 5′ GCCAGGACCAAACCTTTGTGT 3′. PCR was performed using an Eppendorf Mastercycler Nexus GSX1 with conditions as follows: $40\mathrm{ng}$ of genomic DNA was amplified in a total volume of $20\,\mu\mathrm{l}$ containing $0.5\mu\mathrm{mol}$ of each primer, $0.2\mathrm{mmol}$ of each dNTP and 0.02 units of Taq DNA polymerase per reaction. Cycling conditions were 95°C for 3 min, 30 cycles of 95°C for 15 s, 60°C for 30 s, 72°C for 45 s and a final extension step at 72°C for 5 min. PCR products were purified using a QIAquick PCR purification kit. Purified PCR products were capillary sequenced at Source BioScience LifeSciences Genomic Services.

4.4.4 Sequence alignment and variant calling

Sequences were aligned to the California sea lion mtDNA reference genome (NCBI Reference Sequence NC_008416) (Arnason et al., 2006) using Sequencher DNA Sequence Analysis Software (v5.4.6). Alignment errors were inspected manually and corrected. Variant positions were identified by viewing alignments, as well as by manual assessment of sequence chromatograms using FinchTV (v1.4.0). Variants were only assessed within a 397 bp region of the product, comprising region MT:15490–15886 in NC_008416.

4.5 Results

Tumours derived from clonally transmissible cancers carry the genetic material of the original animal that first gave rise to the cancer; thus, transmissible cancers are characterised by shared genotypes that are distinct from those of their matched hosts. To determine whether UGC is clonally transmissible, we assessed 397 base pairs of the mtDNA control region in seven UGC tumours and their matched hosts. The analysis

Table 4.1 Mitochondrial DNA (mtDNA) genotypes at nine polymorphic sites in sea lion hosts and matched tumours. Coordinates are relative to the sea lion mtDNA reference genome, NC_00841621. Individual California sea lions are labelled numerically and matched hosts and tumours are represented side-by-side. Alleles that differ from the reference are shaded in grey.

mtDNA base position	Reference	CSL1 (H)	CSL1 (T)	CSL2 (H)	CSL2 (T)	CSL3 (H)	CSL3 (T)	CSL4 (H)	CSL4 (T)	CSL5 (H)	CSL5 (T)	CSL6 (H)	CSL6 (T)	CSL7 (H)	CSL7 (T)
MT:15524	T	T	T	T	T	T	T	T	T	T	T	C	C	T	T
MT:15527	T	T	T	T	T	T	T	C	C	C	C	C	C	T	T
MT:15528	T	T	T	T	T	T	T	C	C	C	C	C	C	T	T
MT:15550	G	G	G	G	G	G	G	A	A	A	A	A	A	G	G
MT:15551	A	A	A	A	A	A	A	G	G	G	G	A	A	A	A
MT:15629	C	C	C	C	C	C	C	C	C	C	C	C	C	T	T
MT:15652	A	A	A	A	A	A	A	A	A	A	A	G	G	A	A
MT:15660	T	T	T	T	T	T	T	C	C	C	C	C	C	T	T
MT:15812	G	G	G	G	G	G	G	A	A	A	A	A	A	G	G

identified nine polymorphic sites characterising four unique genotypes within the sampled sea lion population (Table 4.1). In all cases, the alleles present in tumours were identical to those found in matched host tissue (Table 4.1). Chromatograms were closely examined at polymorphic sites to rule out the presence of a possibly under-amplified sequence, but no evidence for amplification of additional alleles in tumour tissues was found (Riquet et al., 2017).

4.6 Discussion

This is one of the first systematic studies that attempts to test the hypothesis of clonal transmission of wildlife cancers. Given that transmissible cancers are clonal lineages, tumour cell morphology and tissue architecture is generally very similar between tumours (Karlson and Mann, 1952; Loh et al., 2006). However, previous research has shown that UGCs appear to develop through histologically distinct stages (Lipscomb et al., 2000), which further supports the idea of step-wise oncogenic transformation of host tissue rather than direct transmission of a cancer lineage.

While the results of this study do not support the hypothesis that UGC is clonally transmitted, we cannot exclude the possibility that undetected UGCs may be clonally transmitted. The genital localisation of this cancer, and the likely accessibility of UGC cancer cells to other individuals during coitus, mean that UGC tumours may pose a particular risk for the emergence of a transmissible cancer clone. It is possible

that short-lived transmissible cancers may arise in a population and infect secondary hosts but die out before they are detected.

Future research exploring the role of viral agents, host genetics and environmental factors, as well as somatic genetics, will be important for understanding the carcinogenic processes that cause UGC. It is interesting to note that an OtHV-1-associated UGC has been reported in a South American fur seal (*Arctocephalus australis*) (Dagleish et al., 2013), indicating that other pinnipeds are susceptible to UGC, and further implicating OtHV-1 as a causative agent. Furthermore, analysis of cytological smears collected from California sea lions in the Gulf of California revealed that transformation of the genital epithelium may be relatively common in this species (Barragán-Vargas et al., 2016).

There are a number of limitations concerning this analysis. We only examined genetic variation at one mtDNA locus. As we have shown, at least one transmissible cancer has been observed to sporadically capture mtDNA from its hosts (Rebbeck et al., 2011); thus, mtDNA may not be the most reliable marker for assessing clonality in transmissible cancers. Horizontal transfer of mitochondrial DNA has been detected only twenty times in a cohort of 640 CTVT tumours (Strakova et al., 2016); thus even if mtDNA capture had occurred, it would not be expected that tumours would genetically match their hosts as frequently as we have observed in UGC. There is also the potential risk of amplifying nuclear mitochondrial DNA segments. Although the occurrence of NuMTs has not yet been reported in California sea lions, NuMTs are a major confounder in a wide variety of mammals (Calabrese et al., 2017), as discussed in Section 2.3.7. While this first screen largely refutes the hypothesis that UGC is clonally transmitted, an increased sample size and the inclusion of nuclear molecular markers would undoubtedly provide more robust evidence for this conclusion.

Wildlife models of cancer can provide novel insights into general mechanisms of cancer development (McAloose and Newton, 2009). Like pathogens and parasites, cancer, especially transmissible cancer, can have a negative impact on host fitness in wildlife populations and is an important, but often overlooked, feature of animal ecosystems (Vittecoq et al., 2013). Furthermore, an understanding of the aetiological factors underlying commonly observed cancers in wildlife is essential for conservation and biomonitoring. In this study, we have found no evidence that UGC, one of

the few “cancer epidemics” in wildlife (Browning et al., 2015; McAloose and Newton, 2009), is clonally transmitted. Ruling out this mode of carcinogenesis is an important step in our understanding of UGC, and paves the way towards further research investigating the processes underlying this aggressive disease in sea lions.

Chapter 5

Conclusions and Perspectives

5.1 Summary of the main findings

This thesis describes two major sequencing studies aimed at understanding aspects of the origin and evolution of CTVT. Together, these studies unravel new insights into the biology of transmissible cancers, as well as canine genetics, and suggest exciting future avenues of research.

The main results of this thesis, motivated by the research questions established out in each chapter, are as follows:

1. Horizontal DNA transfer is frequently observed in CTVT. CTVT has captured mtDNA from its transitory hosts via horizontal DNA transfer at least twenty times over the last two thousand years. Eighteen of these mtDNA horizontal transfer events define distinct CTVT clades, whose locations and timings track the global migration patterns of dogs over two millennia. For instance, we found evidence that CTVT was probably introduced to South America concordant with Spanish colonisation, and that CTVT arrived in Australia subsequent to European contact.

2. Recent repeated horizontal transfer of a specific dog mtDNA haplotype. A comparison of CTVT phylogenies generated from nuclear and mitochondrial data revealed 13 instances of mtDNA horizontal transfer involving mtDNA from the same dog mitochondrial haplogroup (A1d1). This intriguing finding suggests that this haplotype may possess either an adaptive or selfish selective advantage facilitating its propagation and maintenance in the CTVT mitochondrial population.

3. A similar process underlies somatic mtDNA mutation in CTVT and in human cancers. The endogenous, strand-asymmetric signature observed in CTVT mtDNA is most likely linked with mtDNA replication and is similar to the mutational signature active in human cancers and in the germline of several animal species.

4. Negative selection acts to limit the accumulation of gene-disrupting mutations in CTVT mtDNA. Intuitively, we might expect negative selection to operate in cancer, however, evidence for its activity has been minimal, possibly due to heterogeneity in cancer vulnerabilities between tissue types. Our analysis thus implicates functional mtDNA as a potential cancer driver in CTVT, providing strong genetic evidence underlining the importance of mtDNA in CTVT evolution.

5. Ancient and recent DNA recombination activity has reorganised tumour mitochondrial genomes in CTVT. Using phylogenetic techniques, we found evidence for an ancient recombination event in a CTVT clade 1 mitochondrial haplotype found throughout Central and South America. Using PacBio long read sequencing, we were able to phase complete mtDNA molecules, identifying evidence for complex real-time mtDNA recombination in a CTVT tumour. This analysis has revealed a novel genetic mechanism, possibly functioning in DNA repair, which may be widespread but previously undetected in cancer.

6. A pre-contact American dog lineage survives as a transmissible canine cancer. Analysis of sequencing data derived from ancient dog remains sampled in North America showed that the CTVT founder dog belonged to a monophyletic lineage of dogs once present throughout the Americas. This thesis reports the first ancient nuclear genomes from multiple pre-contact dogs.

7. CTVT first arose in a founder dog individual that lived up to 8,225 years ago. I reported the first sequencing data from samples involved in a direct CTVT transmission and used this data to calibrate a CTVT molecular clock. Re-analysis of the timing of CTVT's origin is an important contribution to the field and frames the context of CTVT evolution.

8. Pre-contact dogs in the Americas were not domesticated *in situ*. Instead, they were introduced by humans migrating from Asia ~17,000-13,000 years ago from a source population related to modern Arctic dogs. This lineage was subsequently isolated from Eurasian dogs for at least 9,000 years, most likely until a later introduction of dogs during the Thule expansion.

9. Modern dog populations retain minimal genetic ancestry from pre-contact dogs. In fact, most modern dogs in the Americas trace their ancestry from populations introduced by Europeans. Modern Arctic dogs in North America and Siberia share ancestry with PCD, but are not their direct descendants.

10. Analysis of mitochondrial DNA single-nucleotide polymorphisms indicates that urogenital carcinoma in California sea lions is not clonally transmitted (see Chapters 1, 2, and 5.)

The results of Chapters 2 and 3 unravel new insights into the origin, evolution and global spread of CTVT and suggest exciting future avenues of research while Chapter 4 points towards Both studies have also generated data sets that will benefit the wider research community and contribute to future studies of canine genetics (host mitochondrial sequences in Chapter 2; pre-contact mitochondrial and nuclear genome sequences in Chapter 3). Combining modern and ancient DNA analyses as part of collaborative project, this thesis has shed light on when and where CTVT first arose as well as the enigmatic dog population in which the disease is likely to have emerged. Like many recent ancient human studies, this study combines the resolving power of genomics with the time depth of archaeology to reconstruct patterns of evolutionary history.

5.2 Implications and future perspectives

Overall, the data presented in Chapter 2 indicates an important role for functional mitochondria in maintaining CTVT cell fitness. In this context, it is likely that sporadic capture of host mitochondria may have provided a selective advantage in CTVT. Elucidation of these mechanisms goes some way towards understanding how CTVT has managed to maintain genome stability in evolving tumour cell populations.

Major themes for future research arising from this work would be understanding how defective mtDNA may have altered the metabolic environment in CTVT as well as understanding the pathophysiological significance of mtDNA horizontal transfer and recombination. Understanding the mechanism and triggers of horizontal transfer and, in particular, the reason for the recent, repeated uptake of a specific haplotype (see Section 2.7.3) would be a particularly exciting direction in light of the work presented here. Our analysis has reported the first evidence for mtDNA recombination in cancer; not only does this challenge the dogma that mitochondria in animal genomes is recombination inert but this process could be a widespread DNA repair mechanism in cancer, and may be of therapeutic relevance.

One particularly exciting and unexpected direction of research in the context of this thesis is the history of dogs in the Americas.

- First, a further inquiry into the origins and ancestry of dogs in South America will be an interesting area for future study. During our analysis, we found that a mitogenome derived from 1,000-year old dog remains in Argentina clusters within the PCD group, suggesting that PCD contributed ancestry to pre-contact dogs in South America. However, there is little evidence for PCD in modern South American dog populations, suggesting that, similar to North America, dogs harbouring PCD ancestry in South America were replaced by European dogs after colonial contact.
- Second, although this study detected evidence of introgression from wild North American canids (wolves and coyotes) in pre-contact and modern dog populations this thesis does not address which population or populations of wolves and coyotes, when introgression might have taken place, how many introgression events there were and whether introgressed genes are under selection. While qpGraph modelling (Fig. 3.14) suggests that the direction of this admixture is from wolves to dogs, we cannot rule out gene flow from dogs to wolves. Higher coverage genomes from pre-contact dogs as well as both North American wolves and coyotes are necessary to robustly assess the direction of this admixture and map it across the genome. Similarly, estimating the timing of introgression would be challenging and require many high coverage genomes in order to characterise of the size of admixture blocks in the genome.

CTVT is intrinsically interesting as an evolutionary oddity: a mammalian cell rewired for immortality. Assuming that CTVT's infection cycle is roughly six months and that this has been true since the cancer first emerged, then we can estimate that there have been roughly 16,450 transmissions. Like Lenski's famous long-term evolution experiment (LTEE) with E. coli (Lenski et al., 1991), the extreme lifespan of CTVT, make it a unique model for long-term evolutionary processes operating in cancer over several thousand years. Furthermore, study of CTVT provides an opportunity to probe genetic vulnerabilities in cancer and to identify novel host-tumour interactions.

List of Figures

List of Tables

Bibliography

Acevedo-Whitehouse, K., Gulland, F., Greig, D., and Amos, W. (2003). Inbreeding: Disease susceptibility in california sea lions. *Nature*, 422(6927):35.

Agogino, G. A. and Frankforter, W. D. (1960). A Paleo-Indian Bison-Kill in northwestern iowa. *Am. Antiq.*, 25(3):414–415.

Ahmad, T., Mukherjee, S., Pattnaik, B., Kumar, M., Singh, S., Kumar, M., Rehman, R., Tiwari, B. K., Jha, K. A., Barhanpurkar, A. P., Wani, M. R., Roy, S. S., Mabalirajan, U., Ghosh, B., and Agrawal, A. (2014). Miro1 regulates intercellular mitochondrial transport & enhances mesenchymal stem cell rescue efficacy. *EMBO J.*, 33(9):994–1010.

Ajello, P. (1980). Sticker's tumour or transmissible venereal sarcoma. a review, with reference to 590 cases seen at messina, italy since 1931. *Annali della Facolta di Medicina Veterinaria, Messina*, 17:289–339.

Albanese, F., Poli, A., Millanta, F., and Abramo, F. (2002). Primary cutaneous extragenital canine transmissible venereal tumour with leishmania-laden neoplastic cells: a further suggestion of histiocytic origin? *Vet. Dermatol.*, 13(5):243–246.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, 19(9):1655–1664.

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., and Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nat. Genet.*, 47(12):1402–1407.

Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., Campbell, P. J., Vineis, P., Phillips, D. H., and Stratton, M. R. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R.,

van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Zucman-Rossi, J., Futreal, P. A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., and Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421.

Algarra, I., Gaforio, J. J., Cabrera, T., Collado, A., and Garrido, F. (1999). The biological consequences of altered MHC class I expression in tumours. *J. Biol. Regul. Homeost. Agents*, 13(2):90–96.

Allen, G. M. (1920). *Dogs of the American aborigines*, volume 63. Museum of Comparative Zoology.

Allen, G. M. (1939). Dog skulls from uyak bay, kodiak island. *J. Mammal.*, 20(3):336–340.

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlström, T., Vinner, L., Malaspinas, A.-S., Margaryan, A., Higham, T., Chivall, D., Lynnerup, N., Harvig, L., Baron, J., Della Casa, P., Dąbrowski, P., Duffy, P. R., Ebel, A. V., Epimakhov, A., Frei, K., Furmanek, M., Gralak, T., Gromov, A., Gronkiewicz, S., Grupe, G., Hajdu, T., Jarysz, R., Khartanovich, V., Khokhlov, A., Kiss, V., Kolář, J., Kriiska, A., Lasak, I., Longhi, C., McGlynn, G., Merkevicius, A., Merkyte, I., Metspalu, M., Mkrtchyan, R., Moiseyev, V., Paja, L., Pálfi, G., Pokutta, D., Pospieszny, Ł., Price, T. D., Saag, L., Sablin, M., Shishlina, N., Smrčka, V., Soenov, V. I., Szeverényi, V., Tóth, G., Trifanova, S. V., Varul, L., Vicze, M., Yepiskoposyan, L., Zhitenev, V., Orlando, L., Sicheritz-Pontén, T., Brunak, S., Nielsen, R., Kristiansen, K., and Willerslev, E. (2015). Population genomics of bronze age eurasia. *Nature*, 522(7555):167–172.

Amber, E. I., Henderson, R. A., Adeyanju, J. B., and Gyang, E. O. (1990). Single-drug chemotherapy of canine transmissible venereal tumor with cyclophosphamide, methotrexate, or vincristine. *Journal of Veterinary Internal Medicine*.

Anderson, T. M., vonHoldt, B. M., Candille, S. I., Musiani, M., Greco, C., Stahler, D. R., Smith, D. W., Padhukasahasram, B., Randi, E., Leonard, J. A., Bustamante, C. D., Ostrander, E. A., Tang, H., Wayne, R. K., and Barsh, G. S. (2009). Molecular and evolutionary history of melanism in north american gray wolves. *Science*, 323(5919):1339–1343.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N. (1999). Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.*, 23(2):147.

Arnason, U., Gullberg, A., Janke, A., Kullberg, M., Lehman, N., Petrov, E. A., and Väinölä, R. (2006). Pinniped phylogeny and a new hypothesis for their origin and dispersal. *Mol. Phylogenet. Evol.*, 41(2):345–354.

Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of

transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, 79(2):137–158.

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M. T., Perloski, M., Liberg, O., Arnemo, J. M., Hedhammar, A., and Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, 495(7441):360–364.

Baez-Ortega, A. and Gori, K. (2017). Computational approaches for discovery of mutational signatures in cancer. *Brief. Bioinform.*

Bai, B., Zhao, W.-M., Tang, B.-X., Wang, Y.-Q., Wang, L., Zhang, Z., Yang, H.-C., Liu, Y.-H., Zhu, J.-W., Irwin, D. M., Wang, G.-D., and Zhang, Y.-P. (2015). DoGSD: the dog and wolf genome SNP database. *Nucleic Acids Res.*, 43(Database issue):D777–83.

Banfield, W. G., Woke, P. A., Mackay, C. M., and Cooper, H. L. (1965). Mosquito transmission of a reticulum cell sarcoma of hamsters. *Science*, 148(3674):1239–1240.

Barber, B. J. (2004). Neoplastic diseases of commercially important marine bivalves. *Aquat. Living Resour.*, 17(4):449–466.

Barkalow, Jr, F. S. (1972). Vertebrate remains from archeological sites in the tennessee valley of alabama. *South. Indian Stud.*, 24:3–41.

Barragán-Vargas, C., Montano-Frías, J., Ávila Rosales, G., Godínez-Reyes, C. R., and Acevedo-Whitehouse, K. (2016). Transformation of the genital epithelial tract occurs early in california sea lion development. *R Soc Open Sci*, 3(3):150419.

Becsei-Kilborn, E. (2010). Scientific discovery and scientific reputation: the reception of peyton rous' discovery of the chicken sarcoma virus. *J. Hist. Biol.*, 43(1):111–157.

Bellingham Smith, G. and Washbourn, J. W. (1898). Infective venereal tumours in dogs. *J. Comp. Pathol. Ther.*, 11:41–51.

Bergthorsson, U., Adams, K. L., Thomason, B., and Palmer, J. D. (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature*, 424(6945):197–201.

Berridge, M. V., Dong, L., and Neuzil, J. (2015). Mitochondrial DNA in tumor initiation, progression, and metastasis: Role of horizontal mtDNA transfer. *Cancer Res.*, 75(16):3203–3208.

Berridge, M. V., McConnell, M. J., Grasso, C., Bajzikova, M., Kovarova, J., and Neuzil, J. (2016). Horizontal transfer of mitochondria between mammalian cells: beyond co-culture approaches. *Curr. Opin. Genet. Dev.*, 38:75–82.

Björnerfeldt, S., Webster, M. T., and Vilà, C. (2006). Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome Res.*, 16(8):990–994.

Blaine, D. P. (1810). *A Domestic Treatise on the Diseases of Horses and Dogs*. T. Boosey,, London,.

Blick, J. P. (1988). A preliminary report on the osteometric analysis of some aboriginal dogs (canis familiaris) from weyanoke old town, 44 PG 51, prince george county, virginia. archaeological soc. va. Q. *Bull. Am. Assoc. Hist. Nurs.*, 43.

Bolze, P.-A., Weber, B., Fisher, R. A., Seckl, M. J., and Golfier, F. (2013). First confirmation by genotyping of transplacental choriocarcinoma transmission. *Am. J. Obstet. Gynecol.*, 209(4):e4–6.

Boni, M. F., Posada, D., and Feldman, M. W. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics*, 176(2):1035–1047.

Borgic, Q. L. and Galloy, J. M. (2004). Domesticated dog remains from the janey b. goode site. In *Joint Meetings of the Southeastern Archaeological Conference and the Midwestern Archaeological Conference, St. Louis, Missouri, October 21e23*.

Botigué, L. R., Song, S., Scheu, A., Gopalan, S., Pendleton, A. L., Oetjens, M., Taravella, A. M., Seregély, T., Zeeb-Lanz, A., Arbogast, R.-M., Bobo, D., Daly, K., Unterländer, M., Burger, J., Kidd, J. M., and Veeramah, K. R. (2017). Ancient european dog genomes reveal continuity since the early neolithic. *Nat. Commun.*, 8:16082.

Boveri, T. (1914). Zur frage der entstehung maligner tumoren. *Gustav Fischer*.

Bowen, L., Aldridge, B. M., Delong, R., Melin, S., Buckles, E. L., Gulland, F., Lowenstine, L. J., Stott, J. L., and Johnson, M. L. (2005). An immunogenetic basis for the high prevalence of urogenital cancer in a free-ranging population of california sea lions (zalophus californianus). *Immunogenetics*, 56(11):846–848.

Braun-Parvez, L., Charlin, E., Caillard, S., Ducloux, D., Wolf, P., Rolle, F., Golfier, F., Flicoteaux, H., Bergerat, J.-P., and Moulin, B. (2010). Gestational choriocarcinoma transmission following multiorgan donation. *Am. J. Transplant*, 10(11):2541–2546.

Brereton, J. and Haies, E. (1602). *A Brief and True Relation of the Discoveries of the North Part of Virginia*.

Brindley, D. C. and Banfield, W. G. (1961). A contagious tumor of the hamster. *J. Natl. Cancer Inst.*, 26(4):949–957.

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J. G., and Bustamante, C. D. (2012). PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum. Biol.*, 84(4):343–364.

Brønden, L. B., Nielsen, S. S., Toft, N., and Kristensen, A. T. (2010). Data from the danish veterinary cancer registry on the occurrence and distribution of neoplasms in dogs in denmark. *Vet. Rec.*, 166(19):586–590.

Brown, J. A. and Vierra, R. K. (1983). *What happened in the Middle Archaic?: Introduction to an ecological approach to Koster Site Archaeology*. Academic Press Orlando.

Brown, N. O., Calvert, C., and MacEwen, E. G. (1980). Chemotherapeutic management of transmissible venereal tumors in 30 dogs. *J. Am. Vet. Med. Assoc.*, 176(10 Pt 1):983–986.

Brown, S. K., Darwent, C. M., Wictum, E. J., and Sacks, B. N. (2015). Using multiple markers to elucidate the ancient, historical and modern relationships among north american arctic dog breeds. *Heredity*, 115(6):488–495.

Browning, H. M., Acevedo-Whitehouse, K., Gulland, F. M. D., Hall, A. J., Finlayson, J., Dagleish, M. P., Billington, K. J., Colegrove, K., and Hammond, J. A. (2014). Evidence for a genetic basis of urogenital carcinoma in the wild california sea lion. *Proc. Biol. Sci.*, 281(1796):20140240.

Browning, H. M., Gulland, F. M. D., Hammond, J. A., Colegrove, K. M., and Hall, A. J. (2015). Common cancer in a wild animal: the california sea lion (zalophus californianus) as an emerging model for carcinogenesis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 370(1673).

Buck, J. P. (2000). The archaeology and ethnohistory of the dog in virginia algonquian culture as seen from weyanoke old town. *1*, 31.

Buckles, E. L., Lowenstine, L. J., DeLong, R. L., Melin, S. R., Vittore, R. K., Wong, H.-N., Ross, G. L., St Leger, J. A., Greig, D. J., Duerr, R. S., Gulland, F. M. D., and Stott, J. L. (2007). Age-prevalence of otarine herpesvirus-1, a tumor-associated virus, and possibility of its sexual transmission in california sea lions. *Vet. Microbiol.*, 120(1-2):1–8.

Buckles, E. L., Lowenstine, L. J., Funke, C., Vittore, R. K., Wong, H.-N., St Leger, J. A., Greig, D. J., Duerr, R. S., Gulland, F. M. D., and Stott, J. L. (2006). Otarine herpesvirus-1, not papillomavirus, is associated with endemic tumours in california sea lions (zalophus californianus). *J. Comp. Pathol.*, 135(4):183–189.

Butzer, K. W. (1978). Changing holocene environments at the koster site: A Geo-Archaeological perspective. *Am. Antiq.*, 43(3):408–413.

Cairns, J. (1975). Mutation selection and the natural history of cancer. *Nature*, 255(5505):197–200.

Calabrese, F. M., Balacco, D. L., Preste, R., Diroma, M. A., Forino, R., Ventura, M., and Attimonelli, M. (2017). NumtS colonization in mammalian genomes. *Sci. Rep.*, 7(1):16357.

Callaway, E. (2013). Dog genetics spur scientific spat. *Nature*, 498(7454):282–283.

Calvert, C. A., Leifer, C. E., and MacEwen, E. G. (1982). Vincristine for treatment of transmissible venereal tumor in the dog. *J. Am. Vet. Med. Assoc.*, 181(2):163–164.

Camarós, E., Münzel, S. C., Cueto, M., Rivals, F., and Conard, N. J. (2016). The evolution of paleolithic hominin–carnivore interaction written in teeth: Stories from the swabian jura (germany). *Journal of Archaeological Science: Reports*, 6:798–809.

Carreira, V. S., Ferrari, H. F., Langohr, I. M., Mackenzie, C., Montezzo, L. C., Taira, E., Floeter-Winter, L. M., and Luvizotto, M. C. R. (2014). Leishmania sp. amastigotes identification in canine transmissible venereal tumor. *Case Reports in Veterinary Medicine*, 2014.

Castroviejo-Fisher, S., Skoglund, P., Valadez, R., Vilà, C., and Leonard, J. A. (2011). Vanishing native american dog lineages. *BMC Evol. Biol.*, 11:73.

Catone, G., Marino, G., Poglayen, G., Gramiccia, M., Ludovisi, A., and Zanghì, A. (2003). Canine transmissible venereal tumour parasitized by leishmania infantum. *Vet. Res. Commun.*, 27(7):549–553.

Caulin, A. F. and Maley, C. C. (2011). Peto's paradox: evolution's prescription for cancer prevention. *Trends Ecol. Evol.*, 26(4):175–182.

Cavalier-Smith, T. (1987). Eukaryotes with no mitochondria. *Nature*, 326(6111):332–333.

Chapman, R. W., Chapman, R., Kinnes, I., Randsborg, K., Audouze, F., John Disney Professor of Archaeology Cyprian Broodbank, Renfrew, C., Schlanger, N., Sherratt, A., Taylor, T., Yoffee, o. A. N., Ashmore, W., and Murray, o. A. T. (1981). *The Archaeology of Death*. Cambridge University Press.

Charlesworth, B. and Charlesworth, D. (1997). Rapid fixation of deleterious alleles can be caused by muller's ratchet. *Genet. Res.*, 70(1):63–73.

Chatterjee, A., Mambo, E., and Sidransky, D. (2006). Mitochondrial DNA mutations in human cancer. *Oncogene*, 25(34):4663–4674.

Cheng, X. and Ivessa, A. S. (2010). The migration of mitochondrial DNA fragments to the nucleus affects the chronological aging process of saccharomyces cerevisiae. *Aging Cell*, 9(5):919–923.

Chikweto, A., Kumthekar, S., Larkin, H., Deallie, C., Tiwari, K. P., Sharma, R. N., and Bhaiyat, M. I. (2013). Genital and extragenital canine transmissible venereal tumor in dogs in grenada, west indies. *Open Journal of*.

Cho, Y. M., Kim, J. H., Kim, M., Park, S. J., Koh, S. H., Ahn, H. S., Kang, G. H., Lee, J.-B., Park, K. S., and Lee, H. K. (2012). Mesenchymal stem cells transfer mitochondria to the cells with virtually no mitochondrial function but not with pathogenic mtDNA mutations. *PLoS One*, 7(3):e32778.

Clark, D. W. and Milan, F. A. (1974). *Contributions to the later prehistory of Kodiak Island, Alaska*. National Museums of Canada.

Clark, K. A., Howe, D. K., Gafner, K., Kusuma, D., Ping, S., Estes, S., and Denver, D. R. (2012). Selfish little circles: transmission bias and evolution of large deletion-bearing mitochondrial DNA in caenorhabditis briggsae nematodes. *PLoS One*, 7(7):e41433.

Cockrill, J. M. and Beasley, J. N. (1979). Transmission of transmissible venereal tumor of the dog to the coyote. *Am. J. Vet. Res.*, 40(3):409–410.

Cohen, D. (1985). The canine transmissible venereal tumor: a unique result of tumor progression. *Adv. Cancer Res.*, 43:75–112.

Cohen, D., Shalev, A., and Krup, M. (1984). Lack of β2-Microglobulin on the surface of canine transmissible venereal tumor cells. *J. Natl. Cancer Inst.*, 72(2):395–401.

Cohen, D. and Steel, G. G. (1972). Thymidine labelling studies in a transmissible venereal tumour of the dog. *Br. J. Cancer*, 26(5):413–419.

Colegrove, K. M., Gullanda, F. M. D., Naydan, D. K., and Lowenstine, L. J. (2009). Tumor morphology and immunohistochemical expression of estrogen receptor, progesterone receptor, p53, and ki67 in urogenital carcinomas of california sea lions (zalophus californianus). *Vet. Pathol.*, 46(4):642–655.

Coller, H. A., Khrapko, K., Bodyak, N. D., Nekhaeva, E., Herrero-Jimenez, P., and Thilly, W. G. (2001). High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. *Nat. Genet.*, 28(2):147–150.

Concannon, P., Whaley, S., Lein, D., and Wissler, R. (1983). Canine gestation length: variation related to time of mating and fertile life of sperm. *Am. J. Vet. Res.*, 44(10):1819–1821.

Cook, R. A. (2012). Dogs of war: Potential social institutions of conflict, healing, and death in a fort ancient village. *Am. Antiq.*, 77(3):498–523.

Cooper, H. L., Mackay, C. M., and Banfield, W. G. (1964). Chromosome studies of a contagious reticulum cell sarcoma of the syrian hamster. *J. Natl. Cancer Inst.*, 33(4):691–706.

Cooper, M. D. and Alder, M. N. (2006). The evolution of adaptive immune systems. *Cell*, 124(4):815–822.

Coupland, P., Chandra, T., Quail, M., Reik, W., and Swerdlow, H. (2012). Direct sequencing of small genomes on the pacific biosciences RS without library preparation. *Biotechniques*, 53(6):365–372.

Curry, S. S., Brown, D. R., Gaskin, J. M., Jacobson, E. R., Ehrhart, L. M., Blahak, S., Herbst, L. H., and Klein, P. A. (2000). Persistent infectivity of a disease-associated herpesvirus in green turtles after exposure to seawater. *J. Wildl. Dis.*, 36(4):792–797.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., and Meyer, M. (2013). Complete mitochondrial genome sequence of a middle pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.*, 110(39):15758–15763.

Dagleish, M. P., Barrows, M., Maley, M., Killick, R., Finlayson, J., Goodchild, R., Valentine, A., Saunders, R., Willoughby, K., Smith, K. C., and Stidworthy, M. F. (2013). The first report of otarine herpesvirus-1-associated urogenital carcinoma in a south american fur seal (arctocephalus australis). *J. Comp. Pathol.*, 149(1):119–125.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods*, 9(8):772.

David, A. R. and Zimmerman, M. R. (2010). Cancer: an old disease, a new disease or something in between? *Nat. Rev. Cancer*, 10(10):728–733.

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J., Bottomley, W., Davis, N., Dicks, E., Ewing, R., Floyd, Y., Gray, K., Hall, S., Hawes, R., Hughes, J., Kosmidou, V., Menzies, A., Mould, C., Parker, A., Stevens, C., Watt, S., Hooper, S., Wilson, R., Jayatilake, H., Gusterson, B. A., Cooper, C., Shipley, J., Hargrave, D., Pritchard-Jones, K., Maitland, N., Chenevix-Trench, G., Riggins, G. J., Bigner, D. D., Palmieri, G., Cossu, A., Flanagan, A., Nicholson, A., Ho, J. W. C., Leung, S. Y., Yuen, S. T., Weber, B. L., Seigler, H. F., Darrow, T. L., Paterson, H., Marais, R., Marshall, C. J., Wooster, R., Stratton, M. R., and Futreal, P. A. (2002). Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949–954.

Dayan, T. (1994). Early domesticated dogs of the near east. *J. Archaeol. Sci.*, 21(5):633–640.

Deakin, J. E., Bender, H. S., Pearse, A.-M., Rens, W., O'Brien, P. C. M., Ferguson-Smith, M. A., Cheng, Y., Morris, K., Taylor, R., Stuart, A., Belov, K., Amemiya, C. T., Murchison, E. P., Papenfuss, A. T., and Graves, J. A. M. (2012). Genomic restructuring in the tasmanian devil facial tumour: chromosome painting and gene mapping provide clues to evolution of a transmissible tumour. *PLoS Genet.*, 8(2):e1002483.

Decker, B., Davis, B. W., Rimbault, M., Long, A. H., Karlins, E., Jagannathan, V., Reiman, R., Parker, H. G., Drögemüller, C., Corneveaux, J. J., Chapman, E. S., Trent, J. M., Leeb, T., Huentelman, M. J., Wayne, R. K., Karyadi, D. M., and Ostrander, E. A. (2015). Comparison against 186 canid whole-genome sequences reveals survival strategies of an ancient clonally transmissible canine tumor. *Genome Res.*, 25(11):1646–1655.

Derr, M. (2005). *A dog's history of America: how our best friend explored, conquered, and settled a continent*. Macmillan.

Diaz, F., Bayona-Bafaluy, M. P., Rana, M., Mora, M., Hao, H., and Moraes, C. T. (2002). Human mitochondrial DNA with large deletions repopulates organelles faster than full length genomes under relaxed copy number control. *Nucleic Acids Res.*, 30(21):4626–4633.

Dillehay, T. D. and Collins, M. B. (1988). Early cultural evidence from monte verde in chile. *Nature*, 332:150.

Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. L., Abbott, R. M., Hoog, J., Dooling, D. J., Koboldt, D. C., Schmidt, H., Kalicki, J., Zhang, Q., Chen, L., Lin, L., Wendl, M. C., McMichael, J. F., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Appelbaum, E., Deschryver, K., Davies, S., Guintoli, T., Lin, L., Crowder, R., Tao, Y., Snider, J. E., Smith, S. M., Dukes, A. F., Sanderson, G. E., Pohl, C. S., Delehaunty, K. D., Fronick, C. C., Pape, K. A., Reed, J. S., Robinson, J. S., Hodges, J. S., Schierding, W., Dees, N. D., Shen, D., Locke, D. P., Wiechert, M. E., Eldred, J. M., Peck, J. B.,

Oberkfell, B. J., Lolofie, J. T., Du, F., Hawkins, A. E., O'Laughlin, M. D., Bernard, K. E., Cunningham, M., Elliott, G., Mason, M. D., Thompson, Jr, D. M., Ivanovich, J. L., Goodfellow, P. J., Perou, C. M., Weinstock, G. M., Aft, R., Watson, M., Ley, T. J., Wilson, R. K., and Mardis, E. R. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, 464(7291):999–1005.

Domazet-Lošo, T., Klimovich, A., Anokhin, B., Anton-Erxleben, F., Hamm, M. J., Lange, C., and Bosch, T. C. G. (2014). Naturally occurring tumours in the basal metazoan hydra. *Nat. Commun.*, 5:4222.

Dong, L.-F., Kovarova, J., Bajzikova, M., Bezawork-Geleta, A., Svec, D., Endaya, B., Sachaphibulkij, K., Coelho, A. R., Sebkova, N., Ruzickova, A., Tan, A. S., Kluckova, K., Judasova, K., Zamecnikova, K., Rychtarcikova, Z., Gopalan, V., Andera, L., Sobol, M., Yan, B., Pattnaik, B., Bhatraju, N., Truksa, J., Stopka, P., Hozak, P., Lam, A. K., Sedlacek, R., Oliveira, P. J., Kubista, M., Agrawal, A., Dvorakova-Hortova, K., Rohlena, J., Berridge, M. V., and Neuzil, J. (2017). Horizontal transfer of whole mitochondria restores tumorigenic potential in mitochondrial DNA-deficient cancer cells. *Elife*, 6.

Drögemüller, C., Karlsson, E. K., Hytönen, M. K., Perloski, M., Dolf, G., Sainio, K., Lohi, H., Lindblad-Toh, K., and Leeb, T. (2008). A mutation in hairless dogs implicates FOXI3 in ectodermal development. *Science*, 321(5895):1462.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.*, 22(5):1185–1192.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, 29(8):1969–1973.

Druzhkova, A. S., Thalmann, O., Trifonov, V. A., Leonard, J. A., Vorobieva, N. V., Ovodov, N. D., Graphodatsky, A. S., and Wayne, R. K. (2013). Ancient DNA analysis affirms the canid from altai as a primitive dog. *PLoS One*, 8(3):e57754.

Duesberg, P. H. and Vogt, P. K. (1970). Differences between the ribonucleic acids of transforming and nontransforming avian tumor viruses. *Proceedings of the National*.

Dulbecco, R. (1986). A turning point in cancer research: sequencing the human genome. *Science*, 231(4742):1055–1056.

Duncan, J. R. and Prasse, K. W. (1979). Cytology of canine cutaneous round cell tumors. mast cell tumor, histiocytoma, lymphosarcoma and transmissible venereal tumor. *Vet. Pathol.*, 16(6):673–679.

Dungern, V. (1912). Zur biologie des rundezellensarkom des hundes. *Munch. Med. Wochenschr.*, 5:238.

Dürst, M., Gissmann, L., Ikenberg, H., and zur Hausen, H. (1983). A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. *Proc. Natl. Acad. Sci. U. S. A.*, 80(12):3812–3815.

Ebenesersdóttir, S. S., Sandoval-Velasco, M., Gunnarsdóttir, E. D., Jagadeesan, A., Guðmundsdóttir, V. B., Thordardóttir, E. L., Einarsdóttir, M. S., Moore, K. H. S., Sigurðsson, Á., Magnúsdóttir, D. N., Jónsson, H., Snorradóttir, S., Hovig, E., Møller, P., Kockum, I., Olsson, T., Alfredsson, L., Hansen, T. F., Werge, T., Cavalleri, G. L., Gilbert, E., Lalueza-Fox, C., Walser, 3rd, J. W., Kristjánsdóttir, S., Gopalakrishnan, S., Árnadóttir, L., Magnússon, Ó. Þ., Gilbert, M. T. P., Stefánsson, K., and Helgason, A. (2018). Ancient genomes from iceland reveal the making of a human population. *Science*, 360(6392):1028–1032.

Effron, M., Griner, L., and Benirschke, K. (1977). Nature and rate of neoplasia found in captive wild mammals, birds, and reptiles at necropsy. *J. Natl. Cancer Inst.*, 59(1):185–198.

Eldridge, M., Parker, A., Mueller, C., and Crockford, S. (2014). Archaeological investigations at ya asqalu'i/kaien siding, prince rupert harbour. *Report for the British Columbia Archaeology Branch, Victoria: Government of British Columbia*.

Epstein, R. B. and Bennett, B. T. (1974). Histocompatibility typing and course of canine venereal tumors transplanted into unmodified random dogs. *Cancer Res.*, 34(4):788–793.

Eriksson, A. and Manica, A. (2012). Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc. Natl. Acad. Sci. U. S. A.*, 109(35):13956–13960.

Faltas, B. (2011). Cancer is an ancient disease: the case for better palaeoepidemiological and molecular studies. *Nat. Rev. Cancer*, 11(1):76; author reply 76.

Fan, Z., Silva, P., Gronau, I., Wang, S., Armero, A. S., Schweizer, R. M., Ramirez, O., Pollinger, J., Galaverni, M., Ortega Del-Vecchyo, D., Du, L., Zhang, W., Zhang, Z., Xing, J., Vilà, C., Marques-Bonet, T., Godinho, R., Yue, B., and Wayne, R. K. (2016). Worldwide patterns of genomic variation and admixture in gray wolves. *Genome Res.*, 26(2):163–173.

Feldman, W. H. (1929). So-Called infectious sarcoma of the dog in an unusual anatomic situation. *Am. J. Pathol.*, 5(2):183–195.6.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2015). GLOBOCAN 2012 v1. 0, cancer incidence and mortality worldwide: IARC CancerBase no. 11. lyon, france: International agency for research on cancer; 2013.

Ferreira, A. J., Jaggy, A., Varejão, A. P., Ferreira, M. L., Correia, J. M., Mulas, J. M., Almeida, O., Oliveira, P., and Prada, J. (2000). Brain and ocular metastases from a transmissible venereal tumour in a dog. *J. Small Anim. Pract.*, 41(4):165–168.

Fidler, I. J. (2003). The pathogenesis of cancer metastasis: the'seed and soil'hypothesis revisited. *Nat. Rev. Cancer*, 3(6):453.

Fliss, M. S., Usadel, H., Caballero, O. L., Wu, L., Buta, M. R., Eleff, S. M., Jen, J., and Sidransky, D. (2000). Facile detection of mitochondrial DNA mutations in tumors and bodily fluids. *Science*, 287(5460):2017–2019.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U., and Campbell, P. J. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, 43(Database issue):D805–11.

Forment, J. V., Kaidi, A., and Jackson, S. P. (2012). Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nat. Rev. Cancer*, 12(10):663–670.

Fouad, Y. A. and Aanei, C. (2017). Revisiting the hallmarks of cancer. *Am. J. Cancer Res.*, 7(5):1016–1036.

Fowler, M. L. (1959). *Summary Report of Modoc Rock Shelter, 1952, 1953, 1955, 1956*, volume 8. Illinois State Museum.

Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012). mclust version 4 for r: Normal mixture modeling for Model-Based clustering, classification, and density estimation. 2012. *University of Washington: Seattle*.

Frankforter, W. D. and Agogino, G. A. (1960). The simonsen site: Report for the summer of 1959. *Plains Anthropol.*, 5(10):65–70.

Frantz, L. A. F., Mullin, V. E., Pionnier-Capitan, M., Lebrasseur, O., Ollivier, M., Perri, A., Linderholm, A., Mattiangeli, V., Teasdale, M. D., Dimopoulos, E. A., Tresset, A., Duffraisse, M., McCormick, F., Bartosiewicz, L., Gál, E., Nyerges, É. A., Sablin, M. V., Bréhard, S., Mashkour, M., Bălăşescu, A., Gillet, B., Hughes, S., Chassaing, O., Hitte, C., Vigne, J.-D., Dobney, K., Hänni, C., Bradley, D. G., and Larson, G. (2016). Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*, 352(6290):1228–1231.

Freedman, A. H., Gronau, I., Schweizer, R. M., Ortega-Del Vecchyo, D., Han, E., Silva, P. M., Galaverni, M., Fan, Z., Marx, P., Lorente-Galdos, B., Beale, H., Ramirez, O., Hormozdiari, F., Alkan, C., Vilà, C., Squire, K., Geffen, E., Kusak, J., Boyko, A. R., Parker, H. G., Lee, C., Tadigotla, V., Wilton, A., Siepel, A., Bustamante, C. D., Harkins, T. T., Nelson, S. F., Ostrander, E. A., Marques-Bonet, T., Wayne, R. K., and Novembre, J. (2014). Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet.*, 10(1):e1004016.

Fregel, R., Suárez, N. M., Betancor, E., González, A. M., Cabrera, V. M., and Pestano, J. (2015). Mitochondrial DNA haplogroup phylogeny of the dog: Proposal for a cladistic nomenclature. *Mitochondrion*, 22:75–84.

Futato, E. (2002). Middle and late archaic settlement at the perry site, 1lu25. *Journal of Alabama Archaeology*, 48(2):80–92.

Galov, A., Fabbri, E., Caniglia, R., Arbanasić, H., Lapalombella, S., Flotijančić, T., Bešković, I., Galaverni, M., and Randi, E. (2015). First evidence of hybridization between golden jackal (canis aureus) and domestic dog (canis familiaris) as revealed by genetic markers. royal society open science 2: 150450.

Gantenbein, B., Fet, V., Gantenbein-Ritter, I. A., and Balloux, F. (2005). Evidence for recombination in scorpion mitochondrial DNA (scorpiones: Buthidae). *Proc. Biol. Sci.*, 272(1564):697–704.

Gao, D., Vela, I., Sboner, A., Iaquinta, P. J., Karthaus, W. R., Gopalan, A., Dowling, C., Wanjala, J. N., Undvall, E. A., Arora, V. K., Wongvipat, J., Kossai, M., Ramazanoglu, S., Barboza, L. P., Di, W., Cao, Z., Zhang, Q. F., Sirota, I., Ran, L., MacDonald, T. Y., Beltran, H., Mosquera, J.-M., Touijer, K. A., Scardino, P. T., Laudone, V. P., Curtis, K. R., Rathkopf, D. E., Morris, M. J., Danila, D. C., Slovin, S. F., Solomon, S. B., Eastham, J. A., Chi, P., Carver, B., Rubin, M. A., Scher, H. I., Clevers, H., Sawyers, C. L., and Chen, Y. (2014). Organoid cultures derived from patients with advanced prostate cancer. *Cell*, 159(1):176–187.

Gärtner, H. V., Seidl, C., Luckenbach, C., Schumm, G., Seifried, E., Ritter, H., and Bültmann, B. (1996). Genetic analysis of a sarcoma accidentally transplanted from a patient to a surgeon. *N. Engl. J. Med.*, 335(20):1494–1496.

Gavrilovic, B. B., Andersson, K., and Linde Forsberg, C. (2008). Reproductive patterns in the domestic dog–a retrospective study of the drever breed. *Theriogenology*, 70(5):783–794.

Gerlinger, M., Rowan, A. J., Horswell, S., Math, M., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A., and Swanton, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.*, 366(10):883–892.

Germonpré, M., Lázničková-Galetová, M., Losey, R. J., Räikkönen, J., and Sablin, M. V. (2015). Large canids at the gravettian předmostí site, the czech republic: The mandible. *Quat. Int.*, 359-360:261–279.

Germonpré, M., Lázničková-Galetová, M., and Sablin, M. V. (2012). Palaeolithic dog skulls at the gravettian předmostí site, the czech republic. *J. Archaeol. Sci.*, 39(1):184–202.

Germonpré, M., Sablin, M. V., Stevens, R. E., Hedges, R. E. M., Hofreiter, M., Stiller, M., and Després, V. R. (2009). Fossil dogs and wolves from palaeolithic sites in belgium, the ukraine and russia: osteometry, ancient DNA and stable isotopes. *J. Archaeol. Sci.*, 36(2):473–490.

Gey, G. O., Coffman, W. D., and Kubicek, M. T. (1952). Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. *Cancer Res.*, 12:264–265.

Gibbs, M. J., Armstrong, J. S., and Gibbs, A. J. (2000). Sister-scanning: a monte carlo procedure for assessing signals in recombinant sequences. *Bioinformatics*, 16(7):573–582.

Gimeno, E. J., Massone, A. R., Marino, F. P., and Idiart, J. R. (1995). Intermediate filament expression and lectin histochemical features of canine transmissible venereal tumour. *APMIS*, 103(9):645–650.

Ginel, P. J., Molleda, J. M., Novales, M., Martín, E., Margarito, J. M., and López, R. (1995). Primary transmissible venereal tumour in the nasal cavity of a dog. *Vet. Rec.*, 136(9):222–223.

Goebel, T., Waters, M. R., and O'Rourke, D. H. (2008). The late pleistocene dispersal of modern humans in the americas. *Science*, 319(5869):1497–1502.

Gori, K. C. and Baez-Ortega, A. (2018). sigfit: flexible bayesian inference of mutational signatures. *bioRxiv*, 372896.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. (2010). A draft sequence of the neandertal genome. *Science*, 328(5979):710–722.

Gregory, L. B. (1980). The hatch site: A preliminary report. *Virginia Archeological Society, Quarterly Bulletin*, 34:239–248.

Grimm, D. (2015). Dawn of the dog. *Science*, 348(6232):274–279.

Groves, C. P. (1999). The advantages and disadvantages of being domesticated. *Perspectives in Human Biology*, 4(1):1–12.

Gugel, E. A. and Sanders, M. E. (1986). Needle-stick transmission of human colonic adenocarcinoma. *N. Engl. J. Med.*, 315(23):1487.

Guillermo Paez, J., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F. J., Lindeman, N., Boggon, T. J., Naoki, K., Sasaki, H., Fujii, Y., Eck, M. J., Sellers, W. R., Johnson, B. E., and Meyerson, M. (2004). EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. *Science*, 304(5676):1497–1500.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, 59(3):307–321.

Gulland, F. M., Trupkiewicz, J. G., Spraker, T. R., and Lowenstine, L. J. (1996). Metastatic carcinoma of probable transitional cell origin in 66 free-living california sea lions (zalophus californianus), 1979 to 1994. *J. Wildl. Dis.*, 32(2):250–258.

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L. B., Tubio, J. M. C., Papaemmanuil, E., Brewer, D. S., Kallio, H. M. L., Högnäs, G., Annala, M., Kivinummi, K., Goody, V., Latimer, C., O'Meara, S., Dawson, K. J., Isaacs, W., Emmert-Buck, M. R., Nykter, M., Foster, C., Kote-Jarai, Z., Easton, D., Whitaker, H. C., ICGC Prostate Group, Neal, D. E., Cooper, C. S., Eeles, R. A., Visakorpi, T., Campbell, P. J., McDermott, U., Wedge, D. C., and Bova, G. S. (2015). The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357.

Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., Giuliany, R., Rosner, J., Oloumi, A., Shumansky, K., Chin, S., Turashvili, G., Hirst, M., Caldas, C., Marra, M., Aparicio, S., and Shah, S. (2012). Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome*.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R. G., Hallgren, F., Khartanovich, V., Khokhlov, A., Kunst, M., Kuznetsov, P., Meller, H., Mochalov, O., Moiseyev, V., Nicklisch, N., Pichler, S. L., Risch, R., Rojo Guerra, M. A., Roth, C., Szécsényi-Nagy, A., Wahl, J., Meyer, M., Krause, J., Brown, D., Anthony, D., Cooper, A., Alt, K. W., and Reich, D. (2015). Massive migration from the steppe was a source for Indo-European languages in europe. *Nature*, 522(7555):207–211.

Habano, W., Nakamura, S., and Sugai, T. (1998). Microsatellite instability in the mitochondrial DNA of colorectal carcinomas: evidence for mismatch repair systems in mitochondrial genome. *Oncogene*, 17(15):1931–1937.

Haggard, H. W. and Smith, G. M. (1938). Johannes müller and the modern conception of cancer. *Yale J. Biol. Med.*, 10(5):419.b1–436.

Hajdu, S. I. (2016). Pathfinders in oncology from ancient times to the end of the M iddle a ges. *Cancer*, 122(11):1638–1646.

Hajic, E. R. (1990). *Koster site archaeology I: stratigraphy and landscape evolution*. Center for Amer Archeology Pr.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.

Hansemann, D. (1890). Ueber asymmetrische zelltheilung in epithelkrebsen und deren biologische bedeutung. *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin*, 119(2):299–326.

Harmelin, A., Pinthus, J. H., Katzir, N., Kapon, A., Volcani, Y., Amariglio, E. N., and Rehavi, G. (2001). Use of a murine xenograft model for canine transmissible venereal tumor. *Am. J. Vet. Res.*, 62(6):907–911.

Harriot, T. (1588). *A Briefe and True Report of the New Found Land of Virginia: Sir Walter Raleigh's Colony of MDLXXXV*. Privately print.

Hart, J. P. and Rieth, C. B. (2002). *Northeast Subsistence-settlement Change, AD 700-1300*. New York State Museum, New York State Education Department.

Hawkins, C. E., Baars, C., Hesterman, H., Hocking, G. J., Jones, M. E., Lazenby, B., Mann, D., Mooney, N., Pemberton, D., Pyecroft, S., Restani, M., and Wiersma, J. (2006). Emerging disease and population decline of an island endemic, the tasmanian devil sarcophilus harrisii. *Biol. Conserv.*, 131(2):307–324.

Hayakawa, K., Esposito, E., Wang, X., Terasaki, Y., Liu, Y., Xing, C., Ji, X., and Lo, E. H. (2016). Transfer of mitochondria from astrocytes to neurons after stroke. *Nature*, 535(7613):551–555.

Heizer, R. F. (1956). *Archaeology of the Uyak Site, Kodiak Island, Alaska*, volume 17. University of California Press.

Higgins, D. A. (1966). Observations on the canine transmissible venereal tumour as seen in the bahamas. *Vet. Rec.*, 79(3):67–71.

Hill, F. C. (1972). *A Middle Archaic Dog Burial in Illinois*. Foundation for Illinois Archaeology.

Hoarau, G., Holla, S., Lescasse, R., Stam, W. T., and Olsen, J. L. (2002). Heteroplasmy and evidence for recombination in the mitochondrial control region of the flatfish platichthys flesus. *Mol. Biol. Evol.*, 19(12):2261–2264.

Holmes, J. M. (1981). Measurement of the rate of death of canine transmissible venereal tumour cells transplanted into dogs and nude mice. *Res. Vet. Sci.*, 30(2):248–250.

Hordijk, W. and Gascuel, O. (2005). Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics*, 21(24):4338–4347.

Houart, G. L. (1971). Koster: a stratified archaic site in the illinois valley. *Illinois State Museum reports of investigations*, 22:1–63.

Hrdlička, A. (1944). *The Anthropology of Kodiak Island*. Wistar Institute of Anatomy and Biology.

Hsiao, Y.-W., Liao, K.-W., Chung, T.-F., Liu, C.-H., Hsu, C.-D., and Chu, R.-M. (2008). Interactions of host IL-6 and IFN-γ and cancer-derived TGF-β1 on MHC molecule expression during tumor spontaneous regression. *Cancer Immunol. Immunother.*, 57(7):1091–1104.

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.*, 9(5):1322–1332.

Hulton, P. (1984). *America 1585: The Complete Drawings of John White*. University of North Carolina Press.

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans (2012). Biological agents. volume 100 b. a review of human carcinogens. *IARC Monogr. Eval. Carcinog. Risks Hum.*, 100(Pt B):1–441.

Isoda, T., Ford, A. M., Tomizawa, D., van Delft, F. W., De Castro, D. G., Mitsuiki, N., Score, J., Taki, T., Morio, T., Takagi, M., Saji, H., Greaves, M., and Mizutani, S. (2009). Immunologically silent cancer clone transmission from mother to offspring. *Proc. Natl. Acad. Sci. U. S. A.*, 106(42):17882–17885.

Ivanowski, D. (1892). Ueber die mosaikkrankheit der tabakspflanze. *St Petersb Acad Imp Sci Bul*, 35:67–70.

Jackson, C. (1944). The cytology of the contagious (venereal) tumour of the dog. *Onderstepoort Journal of Veterinary Science and Animal Industry*, 20(1).

Jakobsson, M., Pearce, C., Cronin, T. M., Backman, J., Anderson, L. G., Barrientos, N., Bjork, G., Coxall, H., de Boer, A., Mayer, L. A., Morth, C.-M., Nilsson, J., Rattray, J. E., Stranne, C., Semiletov, I., and O'Regan, M. (2017). Post-glacial flooding of the bering land bridge dated to 11 cal ka BP based on new geophysical and sediment records. *Clim. Past*, 13(8):991.

Johnson, S., Lowenstine, L., Gulland, F., Jang, S., Imai, D., Almy, F., Delong, R., and Gardner, I. (2006). Aerobic bacterial flora of the vagina and prepuce of california sea lions (zalophus californianus) and investigation of associations with urogenital carcinoma. *Vet. Microbiol.*, 114(1-2):94–103.

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., and Orlando, L. (2013). mapdamage2.0: fast approximate bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, 29(13):1682–1684.

Ju, Y. S., Alexandrov, L. B., Gerstung, M., Martincorena, I., Nik-Zainal, S., Ramakrishna, M., Davies, H. R., Papaemmanuil, E., Gundem, G., Shlien, A., Bolli, N., Behjati, S., Tarpey, P. S., Nangalia, J., Massie, C. E., Butler, A. P., Teague, J. W., Vassiliou, G. S., Green, A. R., Du, M.-Q., Unnikrishnan, A., Pimanda, J. E., Teh, B. T., Munshi, N., Greaves, M., Vyas, P., El-Naggar, A. K., Santarius, T., Collins, V. P., Grundy, R., Taylor, J. A., Hayes, D. N., Malkin, D., ICGC Breast Cancer Group, ICGC Chronic Myeloid Disorders Group, ICGC Prostate Cancer Group, Foster, C. S., Warren, A. Y., Whitaker, H. C., Brewer, D., Eeles, R., Cooper, C., Neal, D., Visakorpi, T., Isaacs, W. B., Bova, G. S., Flanagan, A. M., Futreal, P. A., Lynch, A. G., Chinnery, P. F., McDermott, U., Stratton, M. R., and Campbell, P. J. (2014). Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *Elife*, 3.

Ju, Y. S., Tubio, J. M. C., Mifsud, W., Fu, B., Davies, H. R., Ramakrishna, M., Li, Y., Yates, L., Gundem, G., Tarpey, P. S., Behjati, S., Papaemmanuil, E., Martin, S., Fullam, A., Gerstung, M., ICGC Prostate Cancer Working Group, ICGC Bone Cancer Working Group, ICGC Breast Cancer Working Group, Nangalia, J., Green, A. R., Caldas, C., Borg, Å., Tutt, A., Lee, M. T. M., van't Veer, L. J., Tan, B. K. T., Aparicio, S., Span, P. N., Martens, J. W. M., Knappskog, S., Vincent-Salomon, A., Børresen-Dale, A.-L., Eyfjörd, J. E., Myklebost, O., Flanagan, A. M., Foster, C., Neal, D. E., Cooper, C., Eeles, R., Bova, S. G., Lakhani, S. R., Desmedt, C., Thomas, G., Richardson, A. L., Purdie, C. A., Thompson, A. M., McDermott, U., Yang, F., Nik-Zainal, S., Campbell, P. J., and Stratton, M. R. (2015). Frequent somatic transfer

of mitochondrial DNA into the nuclear genome of human cancer cells. *Genome Res.*, 25(6):814–824.

Kane, A. E. and Robinson, C. K. (1988). *Dolores Archaeological Program: Anasazi Communities at Dolores, McPhee Village*. US Department of the Interior, Bureau of Reclamation, Engineering and Research Center.

Karlson, A. G. and Mann, F. C. (1952). The transmissible venereal tumor of dogs: observations on forty generations of experimental transfers. *Ann. N. Y. Acad. Sci.*, 54(6):1197–1213.

Karlsson, E. K. and Lindblad-Toh, K. (2008). Leader of the pack: gene mapping in dogs and other model organisms. *Nat. Rev. Genet.*, 9(9):713–725.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780.

Katzir, N., Rechavi, G., Cohen, J. B., Unger, T., Simoni, F., Segal, S., Cohen, D., and Givol, D. (1985). "retroposon" insertion into the cellular oncogene c-myc in canine transmissible venereal tumor. *Proc. Natl. Acad. Sci. U. S. A.*, 82(4):1054–1058.

Kauffman, H. M., McBride, M. A., Cherikh, W. S., Spain, P. C., and Delmonico, F. L. (2002). Transplant tumor registry: donors with central nervous system tumors1. *Transplantation*, 73(4):579–582.

Kerber, J. E. (1997). Native american treatment of dogs in northeastern north america: Archaeological and ethnohistorical perspectives. *Archaeology of Eastern North America*, 25:81–95.

Kim, K. S., Lee, S. E., Jeong, H. W., and Ha, J. H. (1998). The complete nucleotide sequence of the domestic dog (canis familiaris) mitochondrial genome. *Mol. Phylogenet. Evol.*, 10(2):210–220.

Kim, R. N., Kim, D.-S., Choi, S.-H., Yoon, B.-H., Kang, A., Nam, S.-H., Kim, D.-W., Kim, J.-J., Ha, J.-H., Toyoda, A., Fujiyama, A., Kim, A., Kim, M.-Y., Park, K.-H., Lee, K. S., and Park, H.-S. (2012). Genome analysis of the domestic dog (korean jindo) by massively parallel sequencing. *DNA Res.*, 19(3):275–287.

King, D. P., Hure, M. C., Goldstein, T., Aldridge, B. M., Gulland, F. M. D., Saliki, J. T., Buckles, E. L., Lowenstine, L. J., and Stott, J. L. (2002). Otarine herpesvirus-1: a novel gammaherpesvirus associated with urogenital carcinoma in california sea lions (zalophus californianus). *Vet. Microbiol.*, 86(1-2):131–137.

Kleine, T., Maier, U. G., and Leister, D. (2009). DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annu. Rev. Plant Biol.*, 60:115–138.

Knudson, A. G. (1973). Mutation and human cancer. In Kleln, G., Weinhouse, S., and Haddow, A., editors, *Advances in Cancer Research*, volume 17, pages 317–352. Academic Press.

Knudson, Jr, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.*, 68(4):820–823.

Komnenou, A. T., Thomas, A. L. N., Kyriazis, A. P., Poutahidis, T., and Papazoglou, L. G. (2015). Ocular manifestations of canine transmissible venereal tumour: a retrospective study of 25 cases in greece. *Vet. Rec.*, 176(20):523.

Koop, B. F., Burbidge, M., Byun, A., Rink, U., and Crockford, S. J. (2000). Ancient DNA evidence of a separate origin for north american indigenous dogs. *BAR International Series*, 889:271–286.

Kraytsberg, Y., Schwartz, M., Brown, T. A., Ebralidse, K., Kunz, W. S., Clayton, D. A., Vissing, J., and Khrapko, K. (2004). Recombination of human mitochondrial DNA. *Science*, 304(5673):981.

Kuckelman, K. A. (2013). The archaeology of yellow jacket pueblo (site 5MT5): Excavations at a large community center in southwestern colorado.

Kujawski, M., Jarmuz, M., Rydzanicz, M., Szukala, K., Wierzbicka, M., Grenman, R., Golusinski, W., and Szyfter, K. (2004). Frequent chromosome Y loss in primary, second primary and metastatic squamous cell carcinomas of the head and neck region. *Cancer Lett.*, 208(1):95–101.

Ladoukakis, E. D. and Zouros, E. (2001). Direct evidence for homologous recombination in mussel (mytilus galloprovincialis) mitochondrial DNA. *Mol. Biol. Evol.*, 18(7):1168–1175.

Lallensack, R. (2017). Ancient genomes heat up dog domestication debate. *Nature News*.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2017). PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.*, 34(3):772–773.

Larson, G., Karlsson, E. K., Perri, A., Webster, M. T., Ho, S. Y. W., Peters, J., Stahl, P. W., Piper, P. J., Lingaas, F., Fredholm, M., Comstock, K. E., Modiano, J. F., Schelling, C., Agoulnik, A. I., Leegwater, P. A., Dobney, K., Vigne, J.-D., Vilà, C., Andersson, L., and Lindblad-Toh, K. (2012). Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proc. Natl. Acad. Sci. U. S. A.*, 109(23):8878–8883.

Lecène, P. and Lacassagne, A. (1926). Une observation d'inoculation accidentelle d'une tumeur maligne chez l'homme. *Ann. Anat. Pathol.*, 3:97–112.

Lecoq, H. (2001). Discovery of the first virus, the tobacco mosaic virus: 1892 or 1898? *C. R. Acad. Sci. III*, 324(10):929–933.

Lee, E. J., Merriwether, D. A., Kasparov, A. K., Nikolskiy, P. A., Sotnikova, M. V., Pavlova, E. Y., and Pitulko, V. V. (2015). Ancient DNA analysis of the oldest canid species from the siberian arctic and genetic contribution to the domestic dog. *PLoS One*, 10(5):e0125759.

Lenski, R. E., Rose, M. R., Simpson, S. C., and Tadler, S. C. (1991). Long-Term experimental evolution in escherichia coli. i. adaptation and divergence during 2,000 generations. *Am. Nat.*, 138(6):1315–1341.

Leonard, J. A., Wayne, R. K., Wheeler, J., Valadez, R., Guillén, S., and Vilà, C. (2002). Ancient DNA evidence for old world origin of new world dogs. *Science*, 298(5598):1613–1616.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, 3rd, E. J., Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., DeJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.-W., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.-P., Parker, H. G., Pollinger, J. P., Searle, S. M. J., Sutter, N. B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-Zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., Foley, C., Franke, A., Friedrich, D., Gage, D., Garber, M., Gearin, G., Giannoukos, G., Goode, T., Goyette, A., Graham, J., Grandbois, E., Gyaltsen, K., Hafez, N., Hagopian, D., Hagos, B., Hall, J., Healy, C., Hegarty, R., Honan, T., Horn, A., Houde, N., Hughes, L., Hunnicutt, L., Husby, M., Jester, B., Jones, C., Kamat, A., Kanga, B., Kells, C., Khazanovich, D., Kieu, A. C., Kisner, P., Kumar, M., Lance, K., Landers, T., Lara, M., Lee, W., Leger, J.-P., Lennon, N., Leuper, L., LeVine, S., Liu, J., Liu, X., Lokyitsang, Y., Lokyitsang, T., Lui, A., Macdonald, J., Major, J., Marabella, R., Maru, K., Matthews, C., McDonough, S., Mehta, T., Meldrim, J., Melnikov, A., Meneus, L., Mihalev, A., Mihova, T., Miller, K., Mittelman, R., Mlenga, V., Mulrain, L., Munson, G., Navidi, A., Naylor, J., Nguyen, T., Nguyen, N., Nguyen, C., Nguyen, T., Nicol, R., Norbu, N., Norbu, C., Novod, N., Nyima, T., Olandt, P., O'Neill, B., O'Neill, K., Osman, S., Oyono, L., Patti, C., Perrin, D., Phunkhang, P., Pierre, F., Priest, M., Rachupka, A., Raghuraman, S., Rameau, R., Ray, V., Raymond, C., Rege, F., Rise, C., Rogers, J., Rogov, P., Sahalie, J., Settipalli, S., Sharpe, T., Shea, T., Sheehan, M., Sherpa, N., Shi, J., Shih, D., Sloan, J., Smith, C., Sparrow, T., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Stone,

S., Sykes, S., Tchuinga, P., Tenzing, P., Tesfaye, S., Thoulutsang, D., Thoulutsang, Y., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Venkataraman, V., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Yang, S., Yang, X., Young, G., Yu, Q., Zainoun, J., Zembek, L., Zimmer, A., and Lander, E. S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–819.

Lindgreen, S. (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes*, 5:337.

Lindo, J., Huerta-Sánchez, E., Nakagome, S., Rasmussen, M., Petzelt, B., Mitchell, J., Cybulski, J. S., Willerslev, E., DeGiorgio, M., and Malhi, R. S. (2016). A time transect of exomes from a native american population before and after european contact. *Nat. Commun.*, 7:13175.

Lipe, W. D., Morris, J. N., and Kohler, T. A. (1988). Dolores archaeological program: Anasazi communities at dolores: Grass mesa village. *U.S. Department of the Interior (USDI) Bureau of Reclamation, Engineering and Research Center*.

Lipscomb, T. P., Scott, D. P., Garber, R. L., Krafft, A. E., Tsai, M. M., Lichy, J. H., Taubenberger, J. K., Schulman, F. Y., and Gulland, F. M. (2000). Common metastatic carcinoma of california sea lions (zalophus californianus): evidence of genital origin and association with novel gammaherpesvirus. *Vet. Pathol.*, 37(6):609–617.

Lipscomb, T. P., Scott, D. P., and Schulman, F. Y. (2010). Primary site of sea lion carcinomas. *Vet. Pathol.*, 47(1):185; author reply 186.

Lischer, H. E. L. and Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, 28(2):298–299.

Liu, D. and Uzonna, J. E. (2012). The early interaction of leishmania with macrophages and dendritic cells and its influence on the host immune response. *Front. Cell. Infect. Microbiol.*, 2:83.

Liu, K., Ji, K., Guo, L., Wu, W., Lu, H., Shan, P., and Yan, C. (2014). Mesenchymal stem cells rescue injured endothelial cells in an in vitro ischemia–reperfusion model via tunneling nanotube like structure-mediated mitochondrial transfer. *Microvasc. Res.*, 92:10–18.

Locke, K. B., Yeh, F. J., and Hooper, P. T. (1975). Letter: Transmissible venereal tumour in dogs in australia. *Aust. Vet. J.*, 51(9):449.

Loh, R., Bergfeld, J., Hayes, D., O'hara, A., Pyecroft, S., Raidal, S., and Sharpe, R. (2006). The pathology of devil facial tumor disease (DFTD) in tasmanian devils (sarcophilus harrisii). *Vet. Pathol.*, 43(6):890–895.

Loog, L., Thalmann, O., Sinding, M.-H. S., Schuenemann, V. J., Perri, A., Germonpre, M., Bocherens, H., Witt, K. E., Samaniego Castruita, J. A., Velasco, M. S., Lundstrom, I. K. C., Wales, N., Sonet, G., Frantz, L., Schroeder, H., Budd, J., Jimenez, E.-L., Fedorov, S., Gasparyan, B., Kandel, A. W., Lazničkova-Galetova, M., Napierala,

H., Uerpmann, H.-P., Nikolskiy, P. A., Pavlova, E. Y., Pitulko, V. V., Herzig, K.-H., Malhi, R. S., Willerslev, E., Hansen, A. J., Dobney, K., Gilbert, M. T. P., Krause, J., Larson, G., Eriksson, A., and Manica, A. (2018). Modern wolves trace their origin to a late pleistocene expansion from beringia. *bioRxiv*.

Lope, C. P., Masson, M. A., Hare, T. S., and Kú, P. C. D. (2006). The chronology of mayapan: New radiocarbon evidence. *Ancient Mesoamerica*, 17(2):153–175.

Lunt, D. H. and Hyman, B. C. (1997). Animal mitochondrial DNA recombination. *Nature*, 387(6630):247.

Lynch, M., Koskella, B., and Schaack, S. (2006). Mutation pressure and the evolution of organelle genomic architecture. *Science*, 311(5768):1727–1730.

Lynch, T. J., Bell, D. W., Sordella, R., Gurubhagavatula, S., Okimoto, R. A., Brannigan, B. W., Harris, P. L., Haserlat, S. M., Supko, J. G., Haluska, F. G., Louis, D. N., Christiani, D. C., Settleman, J., and Haber, D. A. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.*, 350(21):2129–2139.

Ma, H. and O'Farrell, P. H. (2015). Selections that isolate recombinant mitochondrial genomes in animals. *Elife*, 4.

Ma, H., Xu, H., and O'Farrell, P. H. (2014). Transmission of mitochondrial mutations and action of purifying selection in drosophila melanogaster. *Nat. Genet.*, 46(4):393–397.

Macintyre, G., Goranova, T. E., De Silva, D., Ennis, D., Piskorz, A. M., Eldridge, M., Sie, D., Lewsley, L.-A., Hanif, A., Wilson, C., Dowson, S., Glasspool, R. M., Lockley, M., Brockbank, E., Montes, A., Walther, A., Sundar, S., Edmondson, R., Hall, G. D., Clamp, A., Gourley, C., Hall, M., Fotopoulou, C., Gabra, H., Paul, J., Supernat, A., Millan, D., Hoyle, A., Bryson, G., Nourse, C., Mincarelli, L., Sanchez, L. N., Ylstra, B., Jimenez-Linan, M., Moore, L., Hofmann, O., Markowetz, F., McNeish, I. A., and Brenton, J. D. (2018). Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.*, 50(9):1262–1270.

Mahrouf-Yorgov, M., Augeul, L., Da Silva, C. C., Jourdan, M., Rigolet, M., Manin, S., Ferrera, R., Ovize, M., Henry, A., Guguin, A., Meningaud, J.-P., Dubois-Randé, J.-L., Motterlini, R., Foresti, R., and Rodriguez, A.-M. (2017). Mesenchymal stem cells sense mitochondria released from damaged cells as danger signals to activate their rescue properties. *Cell Death Differ.*, 24:1224.

Manning, P. J. and Martin, P. D. (1970). Metastasis of canine transmissible venereal tumor to the adenohypophysis. *Pathol. Vet.*, 7(2):148–152.

Margulis, L. (1970). *Origin of eukaryotic cells: Evidence and research implications for a theory of the origin and evolution of microbial, plant and animal cells on the precambrian Earth*. Yale University Press.

Marino, G., Gaglio, G., and Zanghì, A. (2012). Clinicopathological study of canine transmissible venereal tumour in leishmaniotic dogs. *J. Small Anim. Pract.*, 53(6):323–327.

Martin, D. and Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics*, 16(6):562–563.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol*, 1(1):vev003.

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., and Campbell, P. J. (2017). Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041.e21.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H., and Campbell, P. J. (2015). Tumor evolution. high burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886.

Masson, M. A. and Peraza Lope, C. (2008). Animal use at the postclassic maya center of mayapán. *Quat. Int.*, 191(1):170–183.

Masson, M. A. and Peraza Lope, C. (2013). The distribution and diversity of faunal exploitation at mayapán. *The archaeology of Mesoamerican animals*, pages 233–279.

Maynard Smith, J. (1992). Analyzing the mosaic structure of genes. *J. Mol. Evol.*, 34(2):126–129.

McAloose, D. and Newton, A. L. (2009). Wildlife cancer: a conservation perspective. *Nat. Rev. Cancer*, 9(7):517–526.

McCary, B. C. (1995). *Indians in Seventeenth-Century Virginia*. Genealogical Publishing Com.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9):1297–1303.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the ensembl API and SNP effect predictor. *Bioinformatics*, 26(16):2069–2070.

Metzger, M. J., Reinisch, C., Sherry, J., and Goff, S. P. (2015). Horizontal transmission of clonal cancer cells causes leukemia in soft-shell clams. *Cell*, 161(2):255–263.

Metzger, M. J., Villalba, A., Carballal, M. J., Iglesias, D., Sherry, J., Reinisch, C., Muttray, A. F., Baldwin, S. A., and Goff, S. P. (2016). Widespread transmission of independent cancer lineages within multiple bivalve species. *Nature*, 534(7609):705–709.

Mills, W. C. (1906). *Explorations of the Baum prehistoric village site*. FJ Heer.

Misdorp, W. (1996). Veterinary cancer epidemiology. *Vet. Q.*, 18(1):32–36.

Mitchell, P. (2015). Did disease constrain the spread of domestic dogs (canis familiaris) into Sub-Saharan africa? *Azania: Archaeological Research in Africa*, 50(1):92–135.

Mitchell, P. (2017). Disease: A hitherto unexplored constraint on the spread of dogs (canis lupus familiaris) in Pre-Columbian south america. *Journal of World Prehistory*, 30(4):301–349.

Monaghan, G. and Peebles, C. S. (2010). The construction, use, and abandonment of angel site mound a tracing the history of a middle mississippian town through its earthworks. *Am. Antiq.*, 75(4):935–953.

Moore, A. E., Rhoads, C. P., and Southam, C. M. (1957). Homotransplantation of human cell lines. *Science*, 125(3239):158–160.

Moore, C. B. (1915). Aboriginal sites on tennessee river. *Journal of the Academy of Natural Sciences of Philadelphia*, 16:169–428.

Moreno-Mayar, J. V., Potter, B. A., Vinner, L., Steinrücken, M., Rasmussen, S., Terhorst, J., Kamm, J. A., Albrechtsen, A., Malaspinas, A.-S., Sikora, M., Reuther, J. D., Irish, J. D., Malhi, R. S., Orlando, L., Song, Y. S., Nielsen, R., Meltzer, D. J., and Willerslev, E. (2018). Terminal pleistocene alaskan genome reveals first founding population of native americans. *Nature*, 553(7687):203–207.

Morey, D. F. (2006). Burying key evidence: the social bond between dogs and people. *J. Archaeol. Sci.*, 33(2):158–175.

Morey, D. F. and Wiant, M. D. (1992). Early holocene domestic dog burials from the north american midwest. *Curr. Anthropol.*, 33(2):224–229.

Moschoi, R., Imbert, V., Nebout, M., Chiche, J., Mary, D., Prebet, T., Saland, E., Castellano, R., Pouyet, L., Collette, Y., Vey, N., Chabannon, C., Recher, C., Sarry, J.-E., Alcor, D., Peyron, J.-F., and Griessinger, E. (2016). Protective mitochondrial transfer from bone marrow stromal cells to acute myeloid leukemic cells during chemotherapy. *Blood*, 128(2):253–264.

Mozos, E., Méndez, A., Gómez-Villamandos, J. C., Martín De Las Mulas, J., and Pérez, J. (1996). Immunohistochemical characterization of canine transmissible venereal tumor. *Vet. Pathol.*, 33(3):257–263.

Muehlenbachs, A., Bhatnagar, J., Agudelo, C. A., Hidron, A., Eberhard, M. L., Mathison, B. A., Frace, M. A., Ito, A., Metcalfe, M. G., Rollin, D. C., Visvesvara, G. S., Pham, C. D., Jones, T. L., Greer, P. W., Vélez Hoyos, A., Olson, P. D., Diazgranados, L. R., and Zaki, S. R. (2015). Malignant transformation of hymenolepis nana in a human host. *N. Engl. J. Med.*, 373(19):1845–1852.

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutat. Res.*, 106:2–9.

Müller, J. (1838). Über den feineran bau and die forman der krankhauten geschwulste. *Berlin G Reimer*.

Murchison, E. P. (2008). Clonally transmissible cancers in dogs and tasmanian devils. *Oncogene*, 27:S19–S30.

Murchison, E. P., Schulz-Trieglaff, O. B., Ning, Z., Alexandrov, L. B., Bauer, M. J., Fu, B., Hims, M., Ding, Z., Ivakhno, S., Stewart, C., Ng, B. L., Wong, W., Aken, B., White, S., Alsop, A., Becq, J., Bignell, G. R., Cheetham, R. K., Cheng, W., Connor, T. R., Cox, A. J., Feng, Z.-P., Gu, Y., Grocock, R. J., Harris, S. R., Khrebtukova, I., Kingsbury, Z., Kowarsky, M., Kreiss, A., Luo, S., Marshall, J., McBride, D. J., Murray, L., Pearse, A.-M., Raine, K., Rasolonjatovo, I., Shaw, R., Tedder, P., Tregidgo, C., Vilella, A. J., Wedge, D. C., Woods, G. M., Gormley, N., Humphray, S., Schroth, G., Smith, G., Hall, K., Searle, S. M. J., Carter, N. P., Papenfuss, A. T., Futreal, P. A., Campbell, P. J., Yang, F., Bentley, D. R., Evers, D. J., and Stratton, M. R. (2012). Genome sequencing and analysis of the tasmanian devil and its transmissible cancer. *Cell*, 148(4):780–791.

Murchison, E. P., Wedge, D. C., Alexandrov, L. B., Fu, B., Martincorena, I., Ning, Z., Tubio, J. M. C., Werner, E. I., Allen, J., De Nardi, A. B., Donelan, E. M., Marino, G., Fassati, A., Campbell, P. J., Yang, F., Burt, A., Weiss, R. A., and Stratton, M. R. (2014). Transmissible [corrected] dog cancer genome reveals the origin and history of an ancient cell lineage. *Science*, 343(6169):437–440.

Murgia, C., Pritchard, J. K., Kim, S. Y., Fassati, A., and Weiss, R. A. (2006). Clonal origin and evolution of a transmissible cancer. *Cell*, 126(3):477–487.

Murphree, A. L. and Benedict, W. F. (1984). Retinoblastoma: clues to human oncogenesis. *Science*, 223(4640):1028–1033.

Murray, M., James, Z. H., and Martin, W. B. (1969). A study of the cytology and karyotype of the canine transmissible venereal tumour. *Res. Vet. Sci.*, 10(6):565–568.

Nak, D., Misirlioglu, D., Nak, Y., Seyrek-Intas, K., and Tek, H. B. (2004). Transmissible venereal tumor with mammary gland metastase in a bitch. *Vet. Bilimleri Derg*, 20:99–102.

Nak, D., Nak, Y., Cangul, I. T., and Tuna, B. (2005). A clinico-pathological study on the effect of vincristine on transmissible venereal tumour in dogs. *Journal of Veterinary Medicine Series A*, 52(7):366–370.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J., and Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, 472:90.

Nerlich, A. G., Rohrbach, H., Bachmeier, B., and Zink, A. (2006). Malignant tumors in two ancient populations: An approach to historical tumor epidemiology. *Oncol. Rep.*, 16(1):197–202.

Ní Leathlobhair, M., Gulland, F. M. D., and Murchison, E. P. (2017). No evidence for clonal transmission of urogenital carcinoma in california sea lions (zalophus californianus). *Wellcome Open Res*, 2:46.

Ní Leathlobhair, M., Perri, A. R., Irving-Pease, E. K., Witt, K. E., Linderholm, A., Haile, J., Lebrasseur, O., Ameen, C., Blick, J., Boyko, A. R., Brace, S., Cortes, Y. N., Crockford, S. J., Devault, A., Dimopoulos, E. A., Eldridge, M., Enk, J., Gopalakrishnan, S., Gori, K., Grimes, V., Guiry, E., Hansen, A. J., Hulme-Beaman, A., Johnson, J., Kitchen, A., Kasparov, A. K., Kwon, Y.-M., Nikolskiy, P. A., Lope, C. P., Manin, A., Martin, T., Meyer, M., Myers, K. N., Omura, M., Rouillard, J.-M., Pavlova, E. Y., Sciulli, P., Sinding, M.-H. S., Strakova, A., Ivanova, V. V., Widga, C., Willerslev, E., Pitulko, V. V., Barnes, I., Gilbert, M. T. P., Dobney, K. M., Malhi, R. S., Murchison, E. P., Larson, G., and Frantz, L. A. F. (2018). The evolutionary history of dogs in the americas. *Science*, 361(6397):81–85.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K. W., Mudie, L. J., Varela, I., McBride, D. J., Bignell, G. R., Cooke, S. L., Shlien, A., Gamble, J., Whitmore, I., Maddison, M., Tarpey, P. S., Davies, H. R., Papaemmanuil, E., Stephens, P. J., McLaren, S., Butler, A. P., Teague, J. W., Jönsson, G., Garber, J. E., Silver, D., Miron, P., Fatima, A., Boyault, S., Langerød, A., Tutt, A., Martens, J. W. M., Aparicio, S. A. J. R., Borg, Å., Salomon, A. V., Thomas, G., Børresen-Dale, A.-L., Richardson, A. L., Neuberger, M. S., Futreal, P. A., Campbell, P. J., Stratton, M. R., and Breast Cancer Working Group of the International Cancer Genome Consortium (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjærde, O. C., Langerød, A., Ringnér, M., Ahn, S.-M., Boyault, S., Brock, J. E., Broeks, A., Butler, A., Desmedt, C., Dirix, L., Dronov, S., Fatima, A., Foekens, J. A., Gerstung, M., Hooijer, G. K. J., Jang, S. J., Jones, D. R., Kim, H.-Y., King, T. A., Krishnamurthy, S., Lee, H. J., Lee, J.-Y., Li, Y., McLaren, S., Menzies, A., Mustonen, V., O'Meara, S., Pauporté, I., Pivot, X., Purdie, C. A., Raine, K., Ramakrishnan, K., Rodríguez-González, F. G., Romieu, G., Sieuwerts, A. M., Simpson, P. T., Shepherd, R., Stebbings, L., Stefansson, O. A., Teague, J., Tommasi, S., Treilleux, I., Van den Eynden, G. G., Vermeulen, P., Vincent-Salomon, A., Yates, L., Caldas, C., Veer, L. V., Tutt, A., Knappskog, S., Tan, B. K. T., Jonkers, J., Borg, Å., Ueno, N. T., Sotiriou, C., Viari, A., Futreal, P. A., Campbell, P. J., Span, P. N., Van Laere, S., Lakhani, S. R., Eyfjord, J. E., Thompson, A. M., Birney, E., Stunnenberg, H. G., van de Vijver, M. J., Martens, J. W. M., Børresen-Dale, A.-L., Richardson, A. L., Kong, G., Thomas, G., and Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534:47.

Nolan, K. C. (2009). Archaeological survey of the reinhardt tract property through a certified local government (CLG) grant on behalf of the city of columbus in harrison township, pickaway county, ohio, volume i: Survey results. *Report Submitted to the Ohio Historic Preservation Office, Columbus, OH*.

Nolan, K. C. (2011). Distributional survey of the reinhardt site (33PI880), pickaway county, ohio: A strategy for deciphering the community structure of a fort ancient village. *MidCont. J. Archaeol.*, 36(1):105–130.

Nolan, K. C. and Cook, R. A. (2010). An evolutionary model of social change in the middle ohio valley: Was social complexity impossible during the late woodland but mandatory during the late prehistoric? *Journal of Anthropological Archaeology*, 29(1):62–79.

Nolan, K. C. and Sciulli, P. (2014). Rejoinder to sciulli and purcell: Two late prehistoric dogs from the reinhardt site (33PI880), pickaway county, ohio. *Pennsylvania Archaeologist*, 84:65–73.

Novinski, M. A. (1876). Zur frage über die impfung der krebsigen geschwulste. *Zentralbl Med. Wissensch.*, 14:790–791.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28.

Odes, E. J., Randolph-Quinney, P. S., Steyn, M., Throckmorton, Z., Smilg, J. S., Zipfel, B., Augustine, T. N., de Beer, F., Hoffman, J. W., Franklin, R. D., and Berger, L. R. a. (2016). Earliest hominin cancer: 1.7-million-year-old osteosarcoma from swartkrans cave, south africa. *S. Afr. J. Sci.*, 112(7/8).

Oduye, O. O., Ikede, B. O., Esuruoso, G. O., and Akpokodje, J. U. (1973). Metastatic transmissible venereal tumour in dogs. *J. Small Anim. Pract.*, 14(10):625–637.

Ohta, S. (2006). Contribution of somatic mutations in the mitochondrial genome to the development of cancer and tolerance against anticancer drugs. *Oncogene*, 25(34):4768–4776.

Okkens, A. C., Hekerman, T. W., de Vogel, J. W., and van Haaften, B. (1993). Influence of litter size and breed on variation in length of gestation in the dog. *Vet. Q.*, 15(4):160–161.

on the Evaluation of Carcinogenic Risks to Humans, I. W. G. (1994). Schistosomes, liver flukes and helicobacter pylori. 61.

Ortner, D. J. and Aufderheide, A. C. (1991). *Human paleopathology: current syntheses and future options*. Smithsonian Institution Press.

Oshimura, M., Sasaki, M., and Makino, S. (1973). Chromosomal banding patterns in primary and transplanted venereal tumors of the dog. *J. Natl. Cancer Inst.*, 51(4):1197–1203.

Osipov, N. E. and Golubeva, V. A. (1976). Diagnosis and treatment of transmissible sarcoma. *Veterinariia*, (7):97–98.

Ovodov, N. D., Crockford, S. J., Kuzmin, Y. V., Higham, T. F. G., Hodgins, G. W. L., and van der Plicht, J. (2011). A 33,000-year-old incipient dog from the altai mountains of siberia: evidence of the earliest domestication disrupted by the last glacial maximum. *PLoS One*, 6(7):e22821.

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., and Harris, S. R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom*, 2(4):e000056.

Pang, J.-F., Kluetsch, C., Zou, X.-J., Zhang, A.-B., Luo, L.-Y., Angleby, H., Ardalan, A., Ekström, C., Sköllermo, A., Lundeberg, J., Matsumura, S., Leitner, T., Zhang, Y.-P., and Savolainen, P. (2009). mtDNA data indicate a single origin for dogs south of yangtze river, less than 16,300 years ago, from numerous wolves. *Mol. Biol. Evol.*, 26(12):2849–2864.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290.

Park, M.-S., Kim, Y., Kang, M.-S., Oh, S.-Y., Cho, D.-Y., Shin, N.-S., and Kim, D.-Y. (2006). Disseminated transmissible venereal tumor in a dog. *J. Vet. Diagn. Invest.*, 18(1):130–133.

Parker, H. G., Dreger, D. L., Rimbault, M., Davis, B. W., Mullen, A. B., Carpintero-Ramirez, G., and Ostrander, E. A. (2017). Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. *Cell Rep.*, 19(4):697–708.

Parker, H. G., Kim, L. V., Sutter, N. B., Carlson, S., Lorentzen, T. D., Malek, T. B., Johnson, G. S., DeFrance, H. B., Ostrander, E. A., and Kruglyak, L. (2004). Genetic structure of the purebred domestic dog. *Science*, 304(5674):1160–1164.

Parmalee, P. W., Paloumpis, A. A., and Wilson, N. (1972). *Animals utilized by Woodland peoples occupying the Apple Creek site, Illinois*. Illinois State Museum.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. *Genetics*, 192(3):1065–1093.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.*, 2(12):e190.

Pearse, A.-M. and Swift, K. (2006). Allograft theory: transmission of devil facial-tumour disease. *Nature*, 439(7076):549.

Perri, A., Widga, C., Lawler, D., Martin, T., Loebel, T., Farnsworth, K., Kohn, L., and Buenger, B. (2018). New evidence of the earliest domestic dogs in the americas. *bioRxiv*.

Perucho, M., Goldfarb, M., Shimizu, K., Lama, C., Fogh, J., and Wigler, M. (1981). Human-tumor-derived cell lines contain common and different transforming genes. *Cell*, 27(3 Pt 2):467–476.

Pesavento, P. A., Agnew, D., Keel, M. K., and Woolard, K. D. (2018). Cancer in wildlife: patterns of emergence. *Nat. Rev. Cancer*.

Peto, R., Roe, F. J., Lee, P. N., Levy, L., and Clack, J. (1975). Cancer and ageing in mice and men. *Br. J. Cancer*, 32(4):411–426.

Pickrell, J. K. and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.*, 8(11):e1002967.

Pionnier-Capitan, M., Bemilli, C., Bodu, P., Célérier, G., Ferrié, J.-G., Fosse, P., Garcià, M., and Vigne, J.-D. (2011). New evidence for upper palaeolithic small domestic dogs in South-Western europe. *J. Archaeol. Sci.*, 38(9):2123–2140.

Pitulko, V. V. and Kasparov, A. K. (2017). Archaeological dogs from the early holocene zhokhov site in the eastern siberian arctic. *Journal of Archaeological Science: Reports*, 13:491–515.

Plummer, M., de Martel, C., Vignat, J., Ferlay, J., Bray, F., and Franceschi, S. (2016). Global burden of cancers attributable to infections in 2012: a synthetic analysis. *Lancet Glob Health*, 4(9):e609–16.

Pollock, H. E. D. (1962). *Mayapan, Yucatan, Mexico*, volume 619. Carnegie Institution of Washington.

Polyak, K., Li, Y., Zhu, H., Lengauer, C., Willson, J. K., Markowitz, S. D., Trush, M. A., Kinzler, K. W., and Vogelstein, B. (1998). Somatic mutations of the mitochondrial genome in human colorectal tumours. *Nat. Genet.*, 20(3):291–293.

Posada, D. and Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. U. S. A.*, 98(24):13757–13762.

Potter, M. A. and Baby, R. S. (1964). Hopewellian dogs. *Ohio Journal of Science*, 64(1).

Pulciani, S., Santos, E., Lauver, A. V., Long, L. K., Aaronson, S. A., and Barbacid, M. (1982). Oncogenes in solid human tumours. *Nature*, 300(5892):539–542.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575.

Pye, R. J., Pemberton, D., Tovar, C., Tubio, J. M. C., Dun, K. A., Fox, S., Darby, J., Hayes, D., Knowles, G. W., Kreiss, A., Siddle, H. V. T., Swift, K., Lyons, A. B., Murchison, E. P., and Woods, G. M. (2016). A second transmissible cancer in tasmanian devils. *Proc. Natl. Acad. Sci. U. S. A.*, 113(2):374–379.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Rafelski, S. M. (2013). Mitochondrial network morphology: building an integrative, geometrical view. *BMC Biol.*, 11:71.

Raghavan, M., DeGiorgio, M., Albrechtsen, A., Moltke, I., Skoglund, P., Korneliussen, T. S., Grønnow, B., Appelt, M., Gulløv, H. C., Friesen, T. M., Fitzhugh, W., Malmström, H., Rasmussen, S., Olsen, J., Melchior, L., Fuller, B. T., Fahrni, S. M., Stafford, Jr, T., Grimes, V., Renouf, M. A. P., Cybulski, J., Lynnerup, N., Lahr, M. M., Britton, K., Knecht, R., Arneborg, J., Metspalu, M., Cornejo, O. E., Malaspinas, A.-S., Wang, Y., Rasmussen, M., Raghavan, V., Hansen, T. V. O., Khusnutdinova, E., Pierre, T., Dneprovsky, K., Andreasen, C., Lange, H., Hayes, M. G., Coltrain, J., Spitsyn, V. A.,

Götherström, A., Orlando, L., Kivisild, T., Villems, R., Crawford, M. H., Nielsen, F. C., Dissing, J., Heinemeier, J., Meldgaard, M., Bustamante, C., O'Rourke, D. H., Jakobsson, M., Gilbert, M. T. P., Nielsen, R., and Willerslev, E. (2014). The genetic prehistory of the new world arctic. *Science*, 345(6200):1255832.

Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M. C., Malaspinas, A.-S., Eriksson, A., Moltke, I., Metspalu, M., Homburger, J. R., Wall, J., Cornejo, O. E., Moreno-Mayar, J. V., Korneliussen, T. S., Pierre, T., Rasmussen, M., Campos, P. F., de Barros Damgaard, P., Allentoft, M. E., Lindo, J., Metspalu, E., Rodríguez-Varela, R., Mansilla, J., Henrickson, C., Seguin-Orlando, A., Malmström, H., Stafford, Jr, T., Shringarpure, S. S., Moreno-Estrada, A., Karmin, M., Tambets, K., Bergström, A., Xue, Y., Warmuth, V., Friend, A. D., Singarayer, J., Valdes, P., Balloux, F., Leboreiro, I., Vera, J. L., Rangel-Villalobos, H., Pettener, D., Luiselli, D., Davis, L. G., Heyer, E., Zollikofer, C. P. E., Ponce de León, M. S., Smith, C. I., Grimes, V., Pike, K.-A., Deal, M., Fuller, B. T., Arriaza, B., Standen, V., Luz, M. F., Ricaut, F., Guidon, N., Osipova, L., Voevoda, M. I., Posukh, O. L., Balanovsky, O., Lavryashina, M., Bogunov, Y., Khusnutdinova, E., Gubina, M., Balanovska, E., Fedorova, S., Litvinov, S., Malyarchuk, B., Derenko, M., Mosher, M. J., Archer, D., Cybulski, J., Petzelt, B., Mitchell, J., Worl, R., Norman, P. J., Parham, P., Kemp, B. M., Kivisild, T., Tyler-Smith, C., Sandhu, M. S., Crawford, M., Villems, R., Smith, D. G., Waters, M. R., Goebel, T., Johnson, J. R., Malhi, R. S., Jakobsson, M., Meltzer, D. J., Manica, A., Durbin, R., Bustamante, C. D., Song, Y. S., Nielsen, R., and Willerslev, E. (2015). POPULATION GENETICS. genomic evidence for the pleistocene and recent population history of native americans. *Science*, 349(6250):aab3884.

Randi, E., Hulva, P., Fabbri, E., Galaverni, M., Galov, A., Kusak, J., Bigi, D., Bolfíková, B. Č., Smetanová, M., and Caniglia, R. (2014). Multilocus detection of wolf x dog hybridization in italy, and guidelines for marker selection. *PLoS One*, 9(1):e86409.

Rebbeck, C. A., Leroi, A. M., and Burt, A. (2011). Mitochondrial capture by a transmissible cancer. *Science*, 331(6015):303.

Rebbeck, C. A., Thomas, R., Breen, M., Leroi, A. M., and Burt, A. (2009). Origins and evolution of a transmissible cancer. *Evolution*, 63(9):2340–2349.

Récamier, J. C. A. (1829). *Recherches sur le traitement du cancer: par la compression méthodique simple ou combinée, et sur l'histoire générale de la même maladie*, volume 1:179-83, 185-93, 198. Gabon.

Reddy, E. P., Reynolds, R. K., Santos, E., and Barbacid, M. (1982). A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature*, 300:149.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009). Reconstructing indian population history. *Nature*, 461(7263):489–494.

Renouf, M. A. P. (1993). Palaeoeskimo seal hunters at port au choix, northwestern newfoundland. *Newfoundland and Labrador Studies*, 9(2).

Rezaei, M., Azizi, S., Shahheidaripour, S., and Rostami, S. (2016). Primary oral and nasal transmissible venereal tumor in a mix-breed dog. *Asian Pac. J. Trop. Biomed.*, 6(5):443–445.

Rick, T. C., Erlandson, J. M., Vellanoweth, R. L., and Braje, T. J. (2005). From pleistocene mariners to complex Hunter-Gatherers: The archaeology of the california channel islands. *J World Prehist*, 19(3):169–228.

Rick, T. C., Walker, P. L., Willis, L. M., Noah, A. C., Erlandson, J. M., Vellanoweth, R. L., Braje, T. J., and Kennett, D. J. (2008). Dogs, humans and island ecosystems: the distribution, antiquity and ecology of domestic dogs (canis familiaris) on california's channel islands, USA. *Holocene*, 18(7):1077–1087.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., WGS500 Consortium, Wilkie, A. O. M., McVean, G., and Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.*, 46(8):912–918.

Riquet, F., Simon, A., and Bierne, N. (2017). Weird genotypes? don't discard them, transmissible cancer could be an explanation. *Evol. Appl.*, 10(2):140–145.

Robin, E. D. and Wong, R. (1988). Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. *J. Cell. Physiol.*, 136(3):507–513.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nat. Biotechnol.*, 29(1):24–26.

Robinson, W. S. and Rountree, H. C. (1991). Pocahontas's people: The powhatan indians of virginia through four centuries. *The Journal of American History*, 78(3):1046.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Rothschild, B. M., Witzke, B. J., and Hershkovitz, I. (1999). Metastatic cancer in the jurassic. *Lancet*, 354(9176):398.

Rous, P. (1911). A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *J. Exp. Med.*, 13(4):397–411.

Rowley, J. D. (1973). Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243(5405):290–293.

Rust, J. H. (1949). Transmissible lymphosarcoma in the dog. *J. Am. Vet. Med. Assoc.*, 114(862):10–14.

Rustom, A., Saffrich, R., Markovic, I., Walther, P., and Gerdes, H.-H. (2004). Nanotubular highways for intercellular organelle transport. *Science*, 303(5660):1007–1010.

Sala-Torra, O., Hanna, C., Loken, M. R., Flowers, M. E. D., Maris, M., Ladne, P. A., Mason, J. R., Senitzer, D., Rodriguez, R., Forman, S. J., Deeg, H. J., and Radich, J. P. (2006). Evidence of donor-derived hematologic malignancies after hematopoietic stem cell transplantation. *Biol. Blood Marrow Transplant.*, 12(5):511–517.

Samso, A. (1965). *Recherches experimentales sur le sarcome de Sticker*. PhD thesis, Thesis, Univ. Paris.

Samuels, Y., Wang, Z., Bardelli, A., Silliman, N., Ptak, J., Szabo, S., Yan, H., Gazdar, A., Powell, S. M., Riggins, G. J., Willson, J. K. V., Markowitz, S., Kinzler, K. W., Vogelstein, B., and Velculescu, V. E. (2004). High frequency of mutations of the PIK3CA gene in human cancers. *Science*, 304(5670):554.

Sandusky, G. E., Carlton, W. W., and Wightman, K. A. (1987). Diagnostic immunohistochemistry of canine round cell tumors. *Vet. Pathol.*, 24(6):495–499.

Sastry, G. A., Narayana, J. V., Rao, P. R., and Christopher, J. (1965). A case of metastatic venereal tumour in a bitch. *Indian Vet. J.*, 42(9):658–660.

Savolainen, P., Zhang, Y.-P., Luo, J., Lundeberg, J., and Leitner, T. (2002). Genetic evidence for an east asian origin of domestic dogs. *Science*, 298(5598):1610–1613.

Scharrer, B. and Lochhead, M. S. (1950). Tumors in the invertebrates: a review. *Cancer Res.*

Schlumberger, H. G. and Lucke, B. H. E. W. (1948). Tumors of fishes, amphibians, and reptiles. *Cancer Res.*, 8(12):657–753.

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Al-Rasheid, K. A. S., Willerslev, E., Krogh, A., and Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, 13:178.

Scott, G. R. (1992). Affinities of prehistoric and modern kodiak islanders and the question of Kachemak-Koniag biological continuity. *Arctic Anthropol.*, pages 150–166.

Scotto, J. and Bailar, 3rd, J. C. (1969). Rigoni-Stern and medical statistics. a nineteenth-century approach to cancer research. *J. Hist. Med. Allied Sci.*, 24(1):65–75.

Shannon, L. M., Boyko, R. H., Castelhano, M., Corey, E., Hayward, J. J., McLean, C., White, M. E., Abi Said, M., Anita, B. A., Bondjengo, N. I., Calero, J., Galov, A., Hedimbi, M., Imam, B., Khalap, R., Lally, D., Masta, A., Oliveira, K. C., Pérez, L., Randall, J., Tam, N. M., Trujillo-Cornejo, F. J., Valeriano, C., Sutter, N. B., Todhunter, R. J., Bustamante, C. D., and Boyko, A. R. (2015). Genetic structure in village dogs reveals a central asian domestication origin. *Proc. Natl. Acad. Sci. U. S. A.*, 112(44):13639–13644.

Shih, C., Padhy, L. C., Murray, M., and Weinberg, R. A. (1981). Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature*, 290(5803):261–264.

Shimkin, M. B. (1955). MA novinsky: A note on the history of transplantation of tumors. *Cancer*, 8(4):653–655.

Siddle, H. V. and Kaufman, J. (2013). A tale of two tumours: comparison of the immune escape strategies of contagious cancers. *Mol. Immunol.*, 55(2):190–193.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.*, 7:539.

Skloot, R. (2017). *The immortal life of Henrietta Lacks*. Broadway Books.

Skoglund, P., Ersmark, E., Palkopoulou, E., and Dalén, L. (2015). Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Curr. Biol.*, 25(11):1515–1519.

Skoglund, P., Posth, C., Sirak, K., Spriggs, M., Valentin, F., Bedford, S., Clark, G. R., Reepmeyer, C., Petchey, F., Fernandes, D., Fu, Q., Harney, E., Lipson, M., Mallick, S., Novak, M., Rohland, N., Stewardson, K., Abdullah, S., Cox, M. P., Friedlaender, F. R., Friedlaender, J. S., Kivisild, T., Koki, G., Kusuma, P., Merriwether, D. A., Ricaut, F.-X., Wee, J. T. S., Patterson, N., Krause, J., Pinhasi, R., and Reich, D. (2016). Genomic insights into the peopling of the southwest pacific. *Nature*, 538(7626):510–513.

Skoglund, P. and Reich, D. (2016). A genomic view of the peopling of the americas. *Curr. Opin. Genet. Dev.*, 41:27–35.

Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31.

Smith, C. S. (1947). An early site in cayuga county, new york. william a. ritchie. researches and transactions of the new york state archaeological association, lewis h. morgan chapter, vol. 10, no. 1. rochester, n. y., 1945. *Am. Antiq.*, 12(3Part1):195–196.

Smith, J. (1632). *The Generall Historie of Virginia, New-England, and the Summer Isles: With the Names of the Adventurers, Planters, and Governours from Their First Beginning, An. 1584, to this Present 1626 : with the Procedings of Those Severall Colonies and the Accidents that Befell Them in All Their Journyes and Discoveries : Also the Maps and Descriptions of All Those Countryes, Their Commodities, People, Government, Customes, and Religion Yet Knowne : Divided Into Sixe Bookes*. I.D. and I.H.

Smith, J. M. and Szathmary, E. (1997). *The Major Transitions in Evolution*. OUP Oxford.

Spees, J. L., Lee, R. H., and Gregory, C. A. (2016). Mechanisms of mesenchymal stem/stromal cell function. *Stem Cell Res. Ther.*, 7(1):125.

Spees, J. L., Olson, S. D., Whitney, M. J., and Prockop, D. J. (2006). Mitochondrial transfer between cells can rescue aerobic respiration. *Proc. Natl. Acad. Sci. U. S. A.*, 103(5):1283–1288.

Spence, J. A., Holt, P. E., Sayer, P. D., Rottcher, D., and Cooper, J. E. (1978). Metastasis of a transmissible venereal tumour to the pituitary. *J. Small Anim. Pract.*, 19(3):175–184.

Stahl, P. W. (1985). Tropical forest cosmology: The cultural context of the early valdivia occupations at loma alta (ecuador).

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Stammnitz, M. R., Coorens, T. H. H., Gori, K. C., Hayes, D., Fu, B., Wang, J., Martin-Herranz, D. E., Alexandrov, L. B., Baez-Ortega, A., Barthorpe, S., Beck, A., Giordano, F., Knowles, G. W., Kwon, Y. M., Hall, G., Price, S., Pye, R. J., Tubio, J. M. C., Siddle, H. V. T., Sohal, S. S., Woods, G. M., McDermott, U., Yang, F., Garnett, M. J., Ning, Z., and Murchison, E. P. (2018). The origins and vulnerabilities of two transmissible cancers in tasmanian devils. *Cancer Cell*, 33(4):607–619.e15.

Stewart, J. B., Alaei-Mahabadi, B., Sabarinathan, R., Samuelsson, T., Gorodkin, J., Gustafsson, C. M., and Larsson, E. (2015). Simultaneous DNA and RNA mapping of somatic mitochondrial mutations across diverse human cancers. *PLoS Genet.*, 11(6):e1005333.

Stewart, J. B. and Chinnery, P. F. (2015). The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.*, 16(9):530–542.

Stewart, J. B., Freyer, C., Elson, J. L., Wredenberg, A., Cansu, Z., Trifunovic, A., and Larsson, N.-G. (2008). Strong purifying selection in transmission of mammalian mitochondrial DNA. *PLoS Biol.*, 6(1):e10.

Sticker, A. (1906). Transplantables rundzellensarkom des hundes. *Z Krebs-forsch*, 4(2):227–314.

Strakova, A. and Murchison, E. P. (2014). The changing global distribution and prevalence of canine transmissible venereal tumour. *BMC Vet. Res.*, 10:168.

Strakova, A., Ní Leathlobhair, M., Wang, G.-D., Yin, T.-T., Airikkala-Otter, I., Allen, J. L., Allum, K. M., Bansse-Issa, L., Bisson, J. L., Castillo Domracheva, A., de Castro, K. F., Corrigan, A. M., Cran, H. R., Crawford, J. T., Cutter, S. M., Delgadillo Keenan, L., Donelan, E. M., Faramade, I. A., Flores Reynoso, E., Fotopoulou, E., Fruean, S. N., Gallardo-Arrieta, F., Glebova, O., Häfelin Manrique, R. F., Henriques, J. J., Ignatenko, N., Koenig, D., Lanza-Perea, M., Lobetti, R., Lopez Quintana, A. M., Losfelt, T., Marino, G., Martincorena, I., Martínez Castañeda, S., Martínez-López, M. F., Meyer, M., Nakanwagi, B., De Nardi, A. B., Neunzig, W., Nixon, S. J., Onsare, M. M., Ortega-Pacheco, A., Peleteiro, M. C., Pye, R. J., Reece, J. F., Rojas Gutierrez, J., Sadia, H., Schmeling, S. K., Shamanova, O., Ssuna, R. K., Steenland-Smit, A. E., Svitich, A., Thoya Ngoka, I., Vițălaru, B. A., de Vos, A. P., de Vos, J. P., Walkinton, O., Wedge, D. C., Wehrle-Martinez, A. S., van der Wel, M. G., Widdowson, S. A., and Murchison, E. P. (2016). Mitochondrial genetic diversity, selection and recombination in a canine transmissible cancer. *Elife*, 5.

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239):719–724.

Strauss, D. C. and Thomas, J. M. (2010). Transmission of donor melanoma by organ transplantation. *Lancet Oncol.*, 11(8):790–796.

Strong, J. A. (1985). Late woodland dog ceremonialism on long island in comparative and temporal perspective. *The Bulletin of the Journal of the New York State Archaeological Association*, 91:32–38.

Stubbs, E. L. and Furth, J. (1934). Experimental studies on venereal sarcoma of the dog. *Am. J. Pathol.*, 10(2):275–286.3.

Tabin, C. J., Bradley, S. M., Bargmann, C. I., Weinberg, R. A., Papageorge, A. G., Scolnick, E. M., Dhar, R., Lowy, D. R., and Chang, E. H. (1982). Mechanism of activation of a human oncogene. *Nature*, 300(5888):143–149.

Tait, S. W. G. and Green, D. R. (2010). Mitochondria and cell death: outer membrane permeabilization and beyond. *Nat. Rev. Mol. Cell Biol.*, 11(9):621–632.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.*, 30(12):2725–2729.

Tan, A. S., Baty, J. W., Dong, L.-F., Bezawork-Geleta, A., Endaya, B., Goodwin, J., Bajzikova, M., Kovarova, J., Peterka, M., Yan, B., Pesdar, E. A., Sobol, M., Filimonenko, A., Stuart, S., Vondrusova, M., Kluckova, K., Sachaphibulkij, K., Rohlena, J., Hozak, P., Truksa, J., Eccles, D., Haupt, L. M., Griffiths, L. R., Neuzil, J., and Berridge, M. V. (2015). Mitochondrial genome acquisition restores respiratory function and tumorigenic potential of cancer cells without mitochondrial DNA. *Cell Metab.*, 21(1):81–94.

Team, R. (2013). R development core team. *RA Lang Environ Stat Comput*, 55:275–286.

Thalmann, O., Shapiro, B., Cui, P., Schuenemann, V. J., Sawyer, S. K., Greenfield, D. L., Germonpré, M. B., Sablin, M. V., López-Giráldez, F., Domingo-Roura, X., Napierala, H., Uerpmann, H.-P., Loponte, D. M., Acosta, A. A., Giemsch, L., Schmitz, R. W., Worthington, B., Buikstra, J. E., Druzhkova, A., Graphodatsky, A. S., Ovodov, N. D., Wahlberg, N., Freedman, A. H., Schweizer, R. M., Koepfli, K.-P., Leonard, J. A., Meyer, M., Krause, J., Pääbo, S., Green, R. E., and Wayne, R. K. (2013). Complete mitochondrial genomes of ancient canids suggest a european origin of domestic dogs. *Science*, 342(6160):871–874.

Thomas, R., Rebbeck, C., Leroi, A. M., Burt, A., and Breen, M. (2009). Extensive conservation of genomic imbalances in canine transmissible venereal tumors (CTVT) detected by microarray-based CGH analysis. *Chromosome Res.*, 17(7):927–934.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, 14(2):178–192.

Thyagarajan, B., Padua, R. A., and Campbell, C. (1996). Mammalian mitochondria possess homologous DNA recombination activity. *J. Biol. Chem.*, 271(44):27536–27543.

Tito, R. Y., Belknap, 3rd, S. L., Sobolik, K. D., Ingraham, R. C., Cleeland, L. M., and Lewis, Jr, C. M. (2011). Brief communication: DNA from early holocene american dog. *Am. J. Phys. Anthropol.*, 145(4):653–657.

Tolar, J. and Neglia, J. P. (2003). Transplacental and other routes of cancer transmission between individuals. *J. Pediatr. Hematol. Oncol.*, 25(6):430–434.

Trigger, B. G. and Sturtevant, W. C. (1978). *Handbook of North American Indians. 15. Northeast*. Smithsonian Inst.

Tuck, J. A. (1970). An archaic indian cemetery in newfoundland. *Sci. Am.*, 222(6):112–121.

Tuck, J. A. (1971). An archaic cemetery at port au choix, newfoundland. *Am. Antiq.*, 36(3):343–358.

Tuck, J. A. (1976). *Ancient people of Port au Choix: the excavation of an Archaic Indian cemetery in Newfoundland*. [St. John's]: Institute of Social and Economic Research Memorial University of Newfoundland.

Turner, C., Killoran, C., Thomas, N. S. T., Rosenberg, M., Chuzhanova, N. A., Johnston, J., Kemel, Y., Cooper, D. N., and Biesecker, L. G. (2003). Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. *Hum. Genet.*, 112(3):303–309.

Ubelaker, D. H. and Sturtevant, W. C. (2007). *Handbook of North American Indians: Environment, origins, and population*. Smithsonian Institution.

Ujvari, B., Dowton, M., and Madsen, T. (2007). Mitochondrial DNA recombination in a free-ranging australian lizard. *Biol. Lett.*, 3(2):189–192.

Ujvari, B., Gatenby, R. A., and Thomas, F. (2016). Transmissible cancers, are they more common than thought? *Evol. Appl.*, 9(5):633–634.

van Asch, B., Zhang, A.-B., Oskarsson, M. C. R., Klütsch, C. F. C., Amorim, A., and Savolainen, P. (2013). Pre-Columbian origins of native american dog breeds, with only limited replacement by european dogs, confirmed by mtDNA analysis. *Proc. Biol. Sci.*, 280(1766):20131142.

Vander Heiden, M. G., Cantley, L. C., and Thompson, C. B. (2009). Understanding the warburg effect: the metabolic requirements of cell proliferation. *Science*, 324(5930):1029–1033.

Varughese, E. E., Singla, V. K., Ratnakaran, U., and Gandotra, V. K. (2012). Successful management of metastatic transmissible venereal tumour to skin of mammary region. *Reprod. Domest. Anim.*, 47 Suppl 6:366–369.

Vaysse, A., Ratnakumar, A., Derrien, T., Axelsson, E., Rosengren Pielberg, G., Sigurdsson, S., Fall, T., Seppälä, E. H., Hansen, M. S. T., Lawley, C. T., Karlsson, E. K., LUPA Consortium, Bannasch, D., Vilà, C., Lohi, H., Galibert, F., Fredholm, M., Häggström, J., Hedhammar, A., André, C., Lindblad-Toh, K., Hitte, C., and Webster, M. T. (2011). Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet.*, 7(10):e1002316.

Velasco-Villa, A., Mauldin, M. R., Shi, M., Escobar, L. E., Gallardo-Romero, N. F., Damon, I., Olson, V. A., Streicker, D. G., and Emerson, G. (2017). The history of rabies in the western hemisphere. *Antiviral Res.*, 146:221–232.

Verscheure, S., Backeljau, T., and Desmyter, S. (2015). In silico discovery of a nearly complete mitochondrial genome numt in the dog (canis lupus familiaris) nuclear genome. *Genetica*, 143(4):453–458.

Vilà, C., Savolainen, P., Maldonado, J. E., Amorim, I. R., Rice, J. E., Honeycutt, R. L., Crandall, K. A., Lundeberg, J., and Wayne, R. K. (1997). Multiple and ancient origins of the domestic dog. *Science*, 276(5319):1687–1689.

Vittecoq, M., Roche, B., Daoust, S. P., Ducasse, H., Missé, D., Abadie, J., Labrut, S., Renaud, F., Gauthier-Clerc, M., and Thomas, F. (2013). Cancer: a missing link in ecosystem functioning? *Trends Ecol. Evol.*, 28(11):628–635.

Vonholdt, B. M., Pollinger, J. P., Lohmueller, K. E., Han, E., Parker, H. G., Quignon, P., Degenhardt, J. D., Boyko, A. R., Earl, D. A., Auton, A., Reynolds, A., Bryc, K., Brisbin, A., Knowles, J. C., Mosher, D. S., Spady, T. C., Elkahloun, A., Geffen, E., Pilot, M., Jedrzejewski, W., Greco, C., Randi, E., Bannasch, D., Wilton, A., Shearman, J., Musiani, M., Cargill, M., Jones, P. G., Qian, Z., Huang, W., Ding, Z.-L., Zhang, Y.-P., Bustamante, C. D., Ostrander, E. A., Novembre, J., and Wayne, R. K. (2010). Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature*, 464(7290):898–902.

Wade, H. (1908). An experimental investigation of infective sarcoma of the dog, with a consideration of its relationship to cancer. *J. Pathol.*, 12(2):384–425.

Wagner, R. P. (1999). Rudolph virchow and the genetic basis of somatic ecology. *Genetics*, 151(3):917–920.

Wai, T., Ao, A., Zhang, X., Cyr, D., Dufort, D., and Shoubridge, E. A. (2010). The role of mitochondrial DNA copy number in mammalian fertility. *Biol. Reprod.*, 83(1):52–62.

Wallace, D. C. (2010). Colloquium paper: bioenergetics, the origins of complexity, and the ascent of man. *Proc. Natl. Acad. Sci. U. S. A.*, 107 Suppl 2:8947–8953.

Wallis, W. D. (1955). *The Micmac Indians of Eastern Canada*. U of Minnesota Press.

Wang, D., Lloyd, A. H., and Timmis, J. N. (2012). Environmental stress increases the entry of cytoplasmic organellar DNA into the nucleus in plants. *Proc. Natl. Acad. Sci. U. S. A.*, 109(7):2444–2448.

Wang, G.-D., Zhai, W., Yang, H.-C., Fan, R.-X., Cao, X., Zhong, L., Wang, L., Liu, F., Wu, H., Cheng, L.-G., Poyarkov, A. D., Poyarkov, Jr, N. A., Tang, S.-S., Zhao, W.-M., Gao, Y., Lv, X.-M., Irwin, D. M., Savolainen, P., Wu, C.-I., and Zhang, Y.-P. (2013). The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat. Commun.*, 4:1860.

Wang, G.-D., Zhai, W., Yang, H.-C., Wang, L., Zhong, L., Liu, Y.-H., Fan, R.-X., Yin, T.-T., Zhu, C.-L., Poyarkov, A. D., Irwin, D. M., Hytönen, M. K., Lohi, H., Wu, C.-I., Savolainen, P., and Zhang, Y.-P. (2016). Out of southern east asia: the natural history of domestic dogs across the world. *Cell Res.*, 26(1):21–33.

Wang, S., Lewis, C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M. V., Molina, J. A., Gallo, C., Mazzotti, G., Poletti, G., Hill, K., Hurtado, A. M., Labuda, D., Klitz, W., Barrantes, R., Bortolini, M. C., Salzano, F. M., Petzl-Erler, M. L., Tsuneto, L. T., Llop, E., Rothhammer, F., Excoffier, L., Feldman, M. W., Rosenberg, N. A., and Ruiz-Linares, A. (2007). Genetic variation and population structure in native americans. *PLoS Genet.*, 3(11):e185.

Wang, X. and Gerdes, H.-H. (2015). Transfer of mitochondria via tunneling nanotubes rescues apoptotic PC12 cells. *Cell Death Differ.*, 22:1181.

Warburg, O., Negelein, E., and Posener, K. (1924). Versuche an überlebendem carcinomgewebe. *Klin. Wochenschr.*, 3(24):1062–1064.

Warren, D. M. (2004). PhD thesis, Indiana University.

Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738.

Wayne, R. K., Geffen, E., Girman, D. J., Koepfli, K. P., Lau, L. M., and Marshall, C. R. (1997). Molecular systematics of the canidae. *Syst. Biol.*, 46(4):622–653.

Webb, K. M. and Allard, M. W. (2009). Mitochondrial genome DNA analysis of the domestic dog: identifying informative SNPs outside of the control region. *J. Forensic Sci.*, 54(2):275–288.

Webb, W. S. and DeJarnette, D. L. (1948a). The flint river site. *Geological Survey of Alabama, Museum Paper*, 23.

Webb, W. S. and DeJarnette, D. L. (1948b). Little bear creek site. *Geological Survey of Alabama, Museum Paper*, 26.

Webb, W. S. and DeJarnette, D. L. (1948c). The perry site, lu25. *Alabama Museum of Natural History, Tuscaloosa, Alabama*.

Weber, W. T., Nowell, P. C., and Hare, W. C. (1965). Chromosome studies of a transplanted and a primary canine venereal sarcoma. *J. Natl. Cancer Inst.*, 35(3):537–547.

Wehr, V. (1878). Demonstration der durch impfung von hund auf hund erzeugten carcinomknötchen. *Langenbecks Arch. Chir. Suppl. II Verh. Dtsch. Ges. Chir.*, 17:52–53.

West, C. F. and Jarvis, K. N. (2015). Osteometric variation in domestic dogs (canis familiaris) from the kodiak archipelago, alaska. *International Journal of Osteoarchaeology*, 25(3):289–298.

Westermann, B. (2010). Mitochondrial fusion and fission in cell life and death. *Nat. Rev. Mol. Cell Biol.*, 11(12):872–884.

Witt, K. E., Judd, K., Kitchen, A., Grier, C., Kohler, T. A., Ortman, S. G., Kemp, B. M., and Malhi, R. S. (2015). DNA analysis of ancient dogs of the americas: identifying possible founding haplotypes and reconstructing population histories. *J. Hum. Evol.*, 79:105–118.

Wolf, J. B., Tautz, D., and Trillmich, F. (2007). Galápagos and californian sea lions are separate species: Genetic analysis of the genus zalophus and its implications for conservation management. *Front. Zool.*, 4:20.

Yang, T. J. and Jones, J. B. (1973). Canine transmissible venereal sarcoma: transplantation studies in neonatal and adult dogs. *J. Natl. Cancer Inst.*, 51(6):1915–1918.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871.

Ylitalo, G. M., Stein, J. E., Hom, T., Johnson, L. L., Tilbury, K. L., Hall, A. J., Rowles, T., Greig, D., Lowenstine, L. J., and Gulland, F. M. D. (2005). The role of organochlorines in cancer-associated mortality in california sea lions (zalophus californianus). *Mar. Pollut. Bull.*, 50(1):30–39.

Yuan, Y., Ju, Y. S., Kim, Y., Li, J., Wang, Y., Yang, Y., Martincorena, I., Creighton, C., Weinstein, J. N., Xu, Y., Han, L., Kim, H.-L., Nakagawa, H., Park, K., Campbell, P. J., Liang, H., and (2017). Comprehensive molecular characterization of mitochondrial genomes in human cancers. *bioRxiv*.

Zander, A. R., Sarpell, S., Cunningham, B., Hess, A. D., Johnston, D. A., and Epstein, R. B. (1980). Plasmapheresis as immunotherapeutic modality in the treatment of the canine venereal tumor. *Exp. Hematol.*, 8(9):1123–1128.

Zeder, M. A. (2012). Pathways to animal domestication. *Biodiversity in agriculture: Domestication, evolution*.

Zhao, J.-H. and Liu, W. (2016). The complete mitochondrial genome of the simao chinese indigenous dog. *Mitochondrial DNA A DNA Mapp Seq Anal*, 27(1):545–546.

Zong, W.-X., Rabinowitz, J. D., and White, E. (2016). Mitochondria and cancer. *Mol. Cell*, 61(5):667–676.

Zsurka, G., Kraytsberg, Y., Kudina, T., Kornblum, C., Elger, C. E., Khrapko, K., and Kunz, W. S. (2005). Recombination of mitochondrial DNA in skeletal muscle of individuals with multiple mitochondrial DNA heteroplasmy. *Nat. Genet.*, 37(8):873–877.

Appendix 1

Publicly available data associated with this thesis

This appendix contains a summary list of all the data related to this thesis that have been made publicly available.

Chapter 2

Sequencing reads associated with this study are available through the European Nucleotide Archive (ENA) with the study accession number PRJEB13152. Complete mtDNA sequence data associated with this study are available in FASTA format at GenBank (https://www.ncbi.nlm.nih.gov/genbank/) with accession numbers: KU290400 - KU291095.

Supplementary file 2.1: Sample information.
Summary of information available for 449 CTVT tumours and 338 hosts sequenced in this study. Includes data on location, year of collection, CTVT mtDNA clade, tumour and host mtDNA haplotypes, breed, age and sex. DOI: 10.7554/eLife.14552.015

Supplementary file 2.2: Sequencing coverage and tumour cell fraction.
a. Average per-base coverage for whole genome (CanFam3.1) and for mtDNA genome (CanFam3.1; NC_002008). (**b.**) List of 11 CTVT hosts with low coverage mtDNA regions. **c.** Estimated tumour cell fraction for 449 CTVT tumours; tumour cell fraction was estimated by calculating the average VAF for variant substitutions present in tumour but not in matched host for each tumour. DOI: 10.7554/eLife.14552.016

Supplementary file 2.3: Confirmation of CTVT diagnosis.
Quantitative PCR (qPCR) was performed for LINE-MYC, a CTVT-specific rearrange-

ment (Katzir et al., 1985). 'Normalised input' represents the relative LINE-MYC input detected in each sample normalised to ACTB (see Section 2.4.2). Normalised input >0.05 as indicative of presence of LINE-MYC. Sufficient DNA was not available for samples 1380T and 1381T; diagnosis in these cases was performed with histopathology. Sample 2T, marked by an asterisk was grown as a xenograft. DOI: 10.7554/eLife.14552.017

Supplementary file 2.4: Single point somatic and germline substitution variant lists.
a. Total number of substitution variants (n = 1005) identified in 449 CTVT tumours. **b.** CTVT tumour somatic substitutions list (n = 928), including the average VAF value normalised for host contamination (see Section 2.4.9.1). Back mutations are not included on the list. **c.** CTVT tumour conservative somatic substitutions list (n = 835), including the average VAF value normalised for host contamination (see Section 2.4.9.1). Back mutations are not included on the list. **d.** Germline clade defining substitutions lists. Substitutions present in the pool of host substitutions and also shared between all samples within a clade (see Section 2.4.9.1). **e.** Potential somatic substitutions lists. Substitutions not present in the pool of host substitutions, but shared between all samples within a clade (see Section 2.4.9.1). (**f.**) Total number of substitution variants (n = 1152) identified in 338 CTVT host samples and 252 publicly available dog mitochondrial genomes (see Supplementary file 2.8). DOI: 10.7554/eLife.14552.018

Supplementary file 2.5: Summary of small insertions and deletions (indels).
a. Total number of insertions and deletions identified in tumours (n = 27), including the average VAF value normalised for host contamination (see Materials and methods) **b.** Total number of homoplasmic insertions and deletions in CTVT hosts (n = 7), including the average VAF value. DOI: 10.7554/eLife.14552.019

Supplementary file 2.6: Annotation of single point substitutions.
Annotation of individual point substitution mutations in **a.** 449 CTVT tumours (see list Supplementary file 2.4a, excluding back mutations) and **b.** 338 CTVT hosts (see list Supplementary file 2.4f). Annotation was performed using Variant Effect Predictor (McLaren et al., 2010). In cases where a single substitution affects two different genes, the two annotations are shown on different lines. DOI: 10.7554/eLife.14552.020

Supplementary file 2.7: Annotation of insertions and deletions (indels).
Annotation of individual indels **a.** unique to CTVT tumours and **b.** homoplasmic in CTVT hosts. Annotation was performed using Variant Effect Predictor (McLaren et al., 2010). DOI: 10.7554/eLife.14552.021

Supplementary file 2.8: Publicly available mitochondrial dog genomes used in the study.
Summary of Genbank accession numbers and metadata for 252 publicly available dog mitochondrial genomes analysed in this Chapter (see Section 2.4.6.8 and Supplementary file 2.4f). DOI: 10.7554/eLife.14552.022

Supplementary file 2.9: Sample mtDNA haplotype lists.
Full mtDNA haplotype names provided for 449 CTVT samples and 338 host samples. DOI: 10.7554/eLife.14552.025

Supplementary file 2.10: Substitution VAF lists.
Substitutions with corresponding VAF (before normalisation) for each of 449 CTVT tumours and 338 CTVT hosts. DOI: 10.7554/eLife.14552.026

Supplementary file 2.11: Indel VAF lists.
Indels with corresponding VAF (before normalisation) for each of 438 CTVT tumours and 334 CTVT hosts. Samples with very high coverage of the mitochondrial genome were excluded from the indel analysis (see Section 2.4.7). DOI: 10.7554/eLife.14552.027

Supplementary file 2.12: Source data for maximum likelihood phylogenetic tree of CTVT mtDNA (Data Set 1).
Maximum likelihood phylogenetic tree constructed for 1039 samples (449 complete CTVT mitochondrial genomes and 590 complete dog mitochondrial genomes). All sequences are labelled with the corresponding sample identifier, country, breed and haplotype name. The sample identifier for CTVT hosts is the sample name (Supplementary file 2.1), the sample identifier for the publicly available dogs is the Genbank accession number. Scale bar corresponds to base substitutions per site. DOI: https://doi.org/10.7554/eLife.14552.004

Supplementary file 2.13: Source data for maximum likelihood phylogenetic trees

of CTVT clades 1 to 5.
Maximum likelihood phylogenetic trees for CTVT mtDNA in **a.** clade 1 (n = 170) **b.** clade 2 (n = 252) **c.** clade 3 (n = 22) **d.** clade 4 (n = 3) and **e.** clade 5 (n = 2), rooted with haplotypes CTVT1 to CTVT5 respectively, which contain clade-defining germline and potential somatic substitutions specific to each clade. Bootstrap values were calculated from 100 bootstrap replicates and are shown where bootstrap values $\geq$60. Scale bars correspond to base substitutions per site. Clade 5 contains only two tumours, which are identical both to each other and to the CTVT5 haplotype; thus the tree for this clade was created separately and does not have a scale bar. DOI: 10.7554/eLife.14552.005

Supplementary file 2.14: Source data supporting mtDNA recombination in CTVT clade 1.
Maximum likelihood cladograms constructed using clade 1 mtDNA positions **a.** MT:1-5429 and **b.** MT:5430-16176 (see Section 2.6.8). Trees were constructed with 153 clade 1 CTVT mtDNAs rooted with the CTVT1 haplotype, which contains clade 1 clade-defining germline and potential somatic substitutions (see Section 2.4.9.1).Bootstrap values were calculated from 100 bootstrap replicates and are shown where bootstrap values $\geq$60. DOI: 10.7554/eLife.14552.014

Supplementary file 2.15: Summary of back mutations
List of back mutations attributable and non-attributable to a putative recombination event. DOI: 10.7554/eLife.14552.024

Chapter 3

Sequencing reads for the ancient dog data analysed in Chapter 3 were deposited at the ENA with project number PRJEB22026. Sequencing reads for the CTVT genomes sequenced as part of this study were deposited at the European Nucleotide Archive (ENA) with project number PRJEB22148.

The following data are available in the Dryad digital repository:

Supplementary file 3.1: Mitochondrial DNA FASTA file.
FASTA file containing 1166 dog mtDNA genomes analysed in Chapter 3. DOI:

10.5061/dryad.s1k47j4/1

Supplementary file 3.2: NEXUS tree data.
Maximum likelihood tree (RAxML) of 1166 dogs mtDNA genomes analysed in Chapter 3. DOI: 10.5061/dryad.s1k47j4/2

Supplementary file 3.3: Genbank summary.
Summary of Genbank accession numbers and publication sources for 1166 mtDNA genomes analysed in Chapter 3. DOI: 10.5061/dryad.s1k47j4/3

Supplementary file 3.4: Genotype plink file (bed).
Plink file (bed) containing genotype data for 54 dogs. DOI: 10.5061/dryad.s1k47j4/4

Supplementary file 3.5: Genotype plink file (bim).
Plink file (bim) containing genotype data for 54 dogs. DOI: 10.5061/dryad.s1k47j4/5

Supplementary file 3.6: Genotype plink file (fam).
Plink file (fam) containing genotype data for 54 dogs. DOI:DOI: 10.5061/dryad.s1k47j4/6

Supplementary file 3.7: Neighbour-joining tree.
Neighbour-joining tree in Fig. 3.4b. DOI: 10.5061/dryad.s1k47j4/7

Supplementary file 3.8: Nexus file for Bayesian tree.
Nexus file used for producing Fig. 3.6 (MKV model in MrBayes). DOI: 10.5061/dryad.s1k47j4/8

Supplementary file 3.9:
Bayesian tree in Fig. 3.6. DOI: 10.5061/dryad.s1k47j4/9

Chapter 5

Chapter 5 presents analysis of matched tumour and host tissues from seven individual California sea lions affected with urogenital carcinoma. Partial mtDNA sequence data associated with this study are available in GenBank with accession numbers: MF000998 - MF001011.

Appendix 2

List of publications associated with this thesis

This appendix provides complete references for articles containing work presented in this thesis.

Chapter 2

Strakova, A.*, **Ní Leathlobhair, M.***, Wang, G.-D., Yin, T.-T., Airikkala-Otter, I., Allen, J. L.,Allum, K. M., Bansse-Issa, L., Bisson, J. L., Castillo Domracheva, A., de Castro, K. F.,Corrigan, A. M., Cran, H. R., Crawford, J. T., Cutter, S. M., Delgadillo Keenan, L., Donelan,E. M., Faramade, I. A., Flores Reynoso, E., Fotopoulou, E., Fruean, S. N., Gallardo-Arrieta,F., Glebova, O., Häfelin Manrique, R. F., Henriques, J. J., Ignatenko, N., Koenig, D.,Lanza-Perea, M., Lobetti, R., Lopez Quintana, A. M., Losfelt, T., Marino, G., Martincorena,I., Martínez Castañeda, S., Martínez-López, M. F., Meyer, M., Nakanwagi, B., De Nardi,A. B., Neunzig, W., Nixon, S. J., Onsare, M. M., Ortega-Pacheco, A., Peleteiro, M. C.,Pye, R. J., Reece, J. F., Rojas Gutierrez, J., Sadia, H., Schmeling, S. K., Shamanova, O.,Ssuna, R. K., Steenland-Smit, A. E., Svitich, A., Thoya Ngoka, I., Viţălaru, B. A., de Vos,A. P., de Vos, J. P., Walkinton, O., Wedge, D. C., Wehrle-Martinez, A. S., van der Wel,M. G., Widdowson, S. A., and Murchison, E. P. (2016) Mitochondrial genetic diversity, selection and recombination in a canine transmissible cancer. *Elife*, 5.

* equal contribution.

Chapter 3

Ní Leathlobhair, M.*, Perri, A. R.*, Irving-Pease, E. K.*, Witt, K. E.*, Linderholm, A.*, Haile, J.,Lebrasseur, O., Ameen, C., Blick, J., Boyko, A. R., Brace, S., Cortes, Y. N.,

Crockford,S. J., Devault, A., Dimopoulos, E. A., Eldridge, M., Enk, J., Gopalakrishnan, S., Gori, K., Grimes, V., Guiry, E., Hansen, A. J., Hulme-Beaman, A., Johnson, J., Kitchen, A., Kasparov, A. K., Kwon, Y.-M., Nikolskiy, P. A., Lope, C. P., Manin, A., Martin, T., Meyer, M., Myers, K. N., Omura, M., Rouillard, J.-M., Pavlova, E. Y., Sciulli, P., Sinding, M.-H. S., Strakova, A., Ivanova, V. V., Widga, C., Willerslev, E., Pitulko, V. V., Barnes, I., Gilbert, M. T. P., Dobney, K. M., Malhi, R. S., Murchison, E. P. ‡, Larson, G. ‡, and Frantz, L. A. F ‡ (2018) The evolutionary history of dogs in the Americas. *Science*, 361(6397):81–85.

* equal contribution. ‡ corresponding author.

Baez-Ortega, A., Gori*, K., Strakova*, A., Allen, J. L., Allum, K. M., Bansse-Issa, L., Thinlay N. Bhutia, Jocelyn L. Bisson, Cristóbal Briceño, Artemio Castillo Domracheva, Anne M. Corrigan, Hugh R. Cran, Jane T. Crawford, Eric Davis, Karina F. de Castro, Andrigo B. de Nardi, Anna P. de Vos, Laura Delgadillo Keenan, Edward M. Donelan, Adela R. Espinoza Huerta, Ibikunle A. Faramade, Mohammed Fazil, Eleni Fotopoulou, Skye N. Fruean, Fanny Gallardo-Arrieta, Olga Glebova, Pagona G. Gouletsou, Rodrigo F. Häfelin Manrique, Joaquim J. G. P. Henriques, Rodrigo S. Horta, Natalia Ignatenko, Yaghouba Kane, Cathy King, Debbie Koenig, Krupa, A., Kruzeniski, S. J., Young-Mi Kwon, Marta Lanza-Perea, Lazyan, M., Lopez Quintana, A. M., Losfelt, T., Marino, G., Simón Martínez Castañeda, Martínez-López, M. F., Meyer, M., Migneco, E. J., Nakanwagi, B., Neal, K. B., Neunzig, W., **Máire Ní Leathlobhair**, Nixon S. J., Ortega-Pacheco, A., Pedraza-Ordoñez, F., Peleteiro, M. C., Katherine Polak, Ruth J. Pye, John F. Reece, Jose Rojas Gutierrez, Haleema Sadia, Sheila K. Schmeling, Olga Shamanova, Alan G. Sherlock, Maximilian Stammnitz, Audrey E. Steenland-Smit, Lester J. Tapia Martínez, Ismail Thoya Ngoka, Cristian G. Torres, Elizabeth M. Tudor, Mirjam G. van der Wel, Bogdan A. Viţălaru, Sevil A. Vural, Oliver Walkinton, Jinhong Wang, Alvaro S. Wehrle-Martinez, Sophie A. E. Widdowson, Michael R. Stratton, Ludmil B. Alexandrov, Inigo Martincorena, Elizabeth P. Murchison (2018). Somatic evolution and global expansion of an ancient transmissible cancer lineage. *Manuscript under review*.

Chapter 4

Ní Leathlobhair, M., Gulland, F. M. D., Murchison, E. P. M. (2017) No evidence for clonal transmission of urogenital carcinoma in California sea lions (Zalophus californianus). *Wellcome Open Research* 2:46.